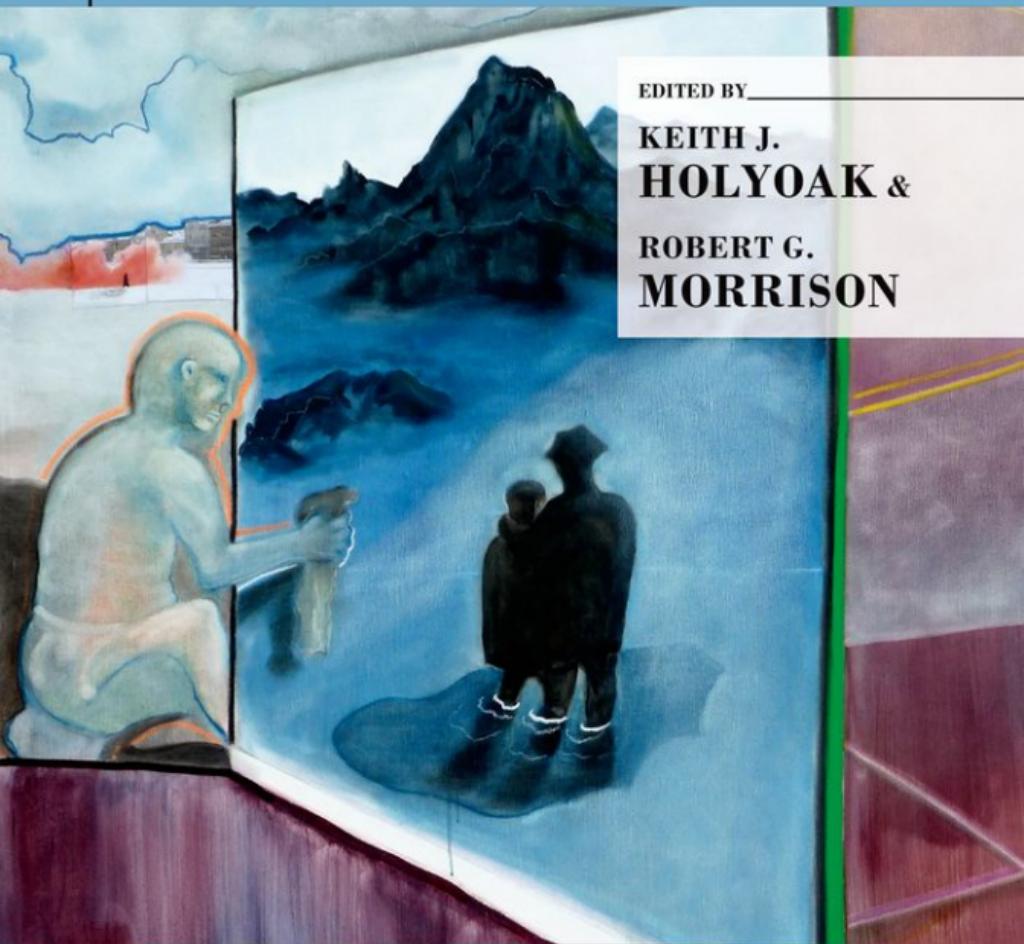




OXFORD LIBRARY OF PSYCHOLOGY

EDITED BY

KEITH J.
HOLYOAK &
ROBERT G.
MORRISON



≡ The Oxford Handbook of
THINKING and
REASONING

The Oxford Handbook of
Thinking and Reasoning

OXFORD LIBRARY OF PSYCHOLOGY

EDITOR-IN-CHIEF

Peter E. Nathan

AREA EDITORS

Clinical Psychology

David H. Barlow

Cognitive Neuroscience

Kevin N. Ochsner and Stephen M. Kosslyn

Cognitive Psychology

Daniel Reisberg

Counseling Psychology

Elizabeth M. Altmaier and Jo-Ida C. Hansen

Developmental Psychology

Philip David Zelazo

Health Psychology

Howard S. Friedman

History of Psychology

David B. Baker

Methods and Measurement

Todd D. Little

Neuropsychology

Kenneth M. Adams

Organizational Psychology

Steve W. J. Kozlowski

Personality and Social Psychology

Kay Deaux and Mark Snyder



OXFORD LIBRARY OF PSYCHOLOGY

Editor in Chief PETER E. NATHAN

The Oxford Handbook of Thinking and Reasoning

Edited by

Keith J. Holyoak

Robert G. Morrison

OXFORD
UNIVERSITY PRESS



Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and certain other countries.

Published in the United States of America by
Oxford University Press
198 Madison Avenue, New York, NY 10016

© Oxford University Press 2012

First issued as an Oxford University Press paperback, 2013.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system,
or transmitted, in any form or by any means, without the prior permission in writing of
Oxford University Press, or as expressly permitted by law, by license, or under terms agreed
with the appropriate reproduction rights organization. Inquiries concerning reproduction
outside the scope of the above should be sent to the Rights Department, Oxford University Press,
at the address above.

You must not circulate this work in any other form
and you must impose this same condition on any acquirer.

Library of Congress Cataloging-in-Publication Data

The Oxford handbook of thinking and reasoning / edited by Keith J. Holyoak, Robert G. Morrison.
p. cm.
ISBN 978-0-19-973468-9 (hardcover); 978-0-19-931379-2 (paperback)
1. Thought and thinking. 2. Reasoning (Psychology) I. Holyoak, Keith James, 1950– II. Morrison,
Robert G., Jr., 1966–
BF441.O94 2012
153.4—dc23
011031592

9 8 7 6 5 4 3 2 1

Printed in the United States of America
on acid-free paper

SHORT CONTENTS

Oxford Library of Psychology vii

About the Editors ix

Preface xi

Contributors xiii

Contents xvii

Chapters 1–806

Index 807

This page intentionally left blank

OXFORD LIBRARY OF PSYCHOLOGY

The *Oxford Library of Psychology*, a landmark series of handbooks, is published by Oxford University Press, one of the world's oldest and most highly respected publishers, with a tradition of publishing significant books in psychology. The ambitious goal of the *Oxford Library of Psychology* is nothing less than to span a vibrant, wide-ranging field and, in so doing, to fill a clear market need.

Encompassing a comprehensive set of handbooks, organized hierarchically, the *Library* incorporates volumes at different levels, each designed to meet a distinct need. At one level are a set of handbooks designed broadly to survey the major subfields of psychology; at another are numerous handbooks that cover important current focal research and scholarly areas of psychology in depth and detail. Planned as a reflection of the dynamism of psychology, the *Library* will grow and expand as psychology itself develops, thereby highlighting significant new research that will impact on the field. Adding to its accessibility and ease of use, the *Library* will be published in print and, later on, electronically.

The *Library* surveys psychology's principal subfields with a set of handbooks that capture the current status and future prospects of those major subdisciplines. This initial set includes handbooks of social and personality psychology, clinical psychology, counseling psychology, school psychology, educational psychology, industrial and organizational psychology, cognitive psychology, cognitive neuroscience, methods and measurements, history, neuropsychology, personality assessment, developmental psychology, and more. Each handbook undertakes to review one of psychology's major subdisciplines with breadth, comprehensiveness, and exemplary scholarship. In addition to these broadly conceived volumes, the *Library* also includes a large number of handbooks designed to explore in depth more specialized areas of scholarship and research, such as stress, health and coping, anxiety and related disorders, cognitive development, or child and adolescent assessment. In contrast to the broad coverage of the subfield handbooks, each of these latter volumes focuses on an especially productive, more highly focused line of scholarship and research. Whether at the broadest or most specific level, however, all of the *Library* handbooks offer synthetic coverage that reviews and evaluates the relevant past and present research and anticipates research in the future. Each handbook in the *Library* includes introductory and concluding chapters written by its editor to provide a roadmap to the handbook's table of contents and to offer informed anticipations of significant future developments in that field.

An undertaking of this scope calls for handbook editors and chapter authors who are established scholars in the areas about which they write. Many of the nation's and world's most productive and best-respected psychologists have

agreed to edit *Library* handbooks or write authoritative chapters in their areas of expertise.

For whom has the *Oxford Library of Psychology* been written? Because of its breadth, depth, and accessibility, the *Library* serves a diverse audience, including graduate students in psychology and their faculty mentors, scholars, researchers, and practitioners in psychology and related fields. These individuals will find in the *Library* the information they seek on the subfield or focal area of psychology in which they work or are interested.

Befitting its commitment to accessibility, each handbook includes a comprehensive index, as well as extensive references to help guide research. And because the *Library* was designed from its inception as an online as well as a print resource, its structure and contents will be readily and rationally searchable online. Furthermore, once the *Library* is released online, the handbooks will be regularly and thoroughly updated.

In summary, the *Oxford Library of Psychology* will grow organically to provide a thoroughly informed perspective on the field of psychology, one that reflects both psychology's dynamism and its increasing interdisciplinarity. Once published electronically, the *Library* is also destined to become a uniquely valuable interactive tool, with extended search and browsing capabilities. As you begin to consult this handbook, we sincerely hope you will share our enthusiasm for the more than 500-year tradition of Oxford University Press for excellence, innovation, and quality, as exemplified by the *Oxford Library of Psychology*.

Peter E. Nathan
Editor-in-Chief
Oxford Library of Psychology

ABOUT THE EDITORS

Keith J. Holyoak

Keith J. Holyoak, Ph.D., Distinguished Professor of Psychology at the University of California, Los Angeles, is a leading researcher in human thinking and reasoning. He received his B.A. from the University of British Columbia in 1971 and his Ph.D. from Stanford University in 1976. Dr. Holyoak was on the faculty of the University of Michigan from 1976–1986 and then joined the Department of Psychology at UCLA. His work combines behavioral studies with both cognitive neuroscience and computational modeling. Dr. Holyoak has published over 180 scientific articles, and is the co-author or editor of numerous books, including *Induction: Processes of Inference, Learning and Discovery* (MIT Press, 1986), *Mental Leaps: Analogy in Creative Thought* (MIT Press, 1995), and the *Cambridge Handbook of Thinking and Reasoning* (Cambridge University Press, 2005). In a parallel career as a poet he has published *Facing the Moon: Poems of Li Bai and Du Fu* (Oyster River Press, 2007), *My Minotaur: Selected Poems 1998–2006* (Dos Madres Press, 2010), and *Foreigner: New English Poems in Chinese Old Style* (Dos Madres Press, 2012).

Robert G. Morrison

Robert G. Morrison, Ph.D., Assistant Professor of Psychology and Neuroscience at Loyola University Chicago, uses behavioral, computational, and neuroimaging methods to investigate memory and reasoning throughout the lifespan. After receiving his B.S. from Wheaton College, Morrison worked both as a scientist and a conceptual artist. He discovered cognitive science at the interface of these worlds and studied Cognitive Neuroscience at UCLA, where he received his Ph.D. in 2004. After co-founding Xunesis (xunesis.org) and completing a National Institute of Aging post-doctoral fellowship at Northwestern University, Morrison joined the Department of Psychology and the Neuroscience Institute at Loyola in 2009. Dr. Morrison has published numerous scientific articles and chapters and has edited the *Cambridge Handbook of Thinking and Reasoning* (Cambridge University Press, 2005). Dr. Morrison's research has been funded by the National Institute of Mental Health, the Office of Naval Research, the American Federation of Aging Research, and the Illinois Department of Public Health. In a parallel career as an artist he has exhibited his painting, sculpture and photography in galleries and museums throughout the United States.

This page intentionally left blank

PREFACE

A few decades ago, when the science of cognition was in its infancy, the early textbooks on cognition began with perception and attention and ended with memory. So-called higher-level cognition—the mysterious, complicated realm of thinking and reasoning—was simply left out. Things changed—any good cognitive text (and there are many) devotes several chapters to topics such as categorization, various types of reasoning, judgment and decision making, and problem solving. As the new century began, we noticed that unlike fields such as perception or memory, the field of thinking and reasoning lacked a true Handbook—a book meant to be kept close “at hand” by those involved in the field, particularly those new to it. In response, we edited the *Cambridge Handbook of Thinking and Reasoning* (2005). Our aim was to bring together top researchers to write chapters each of which summarized the basic concepts and findings for a major topic, sketch its history, and give a sense of the directions in which research is currently heading. The *Handbook* provided quick overviews for experts in each topic area, and more important for experts in allied topic areas (as few researchers can keep up with the scientific literature over the full breadth of the field of thinking and reasoning). Even more crucially, this *Handbook* was meant to provide an entry point into the field for the next generation of researchers, by providing a text for use in classes on thinking and reasoning designed for graduate students and upper-level undergraduates.

The first *Handbook* achieved these aims. However, a fast-moving scientific field has a way of quickly rendering the “state of the art” the “state of yesterday.” By the time the book appeared, new developments that our book had barely touched on were already generating excitement among researchers. These new themes included advances in Bayesian modeling, which helped to understand the rational foundations of thinking and reasoning, and advances in cognitive neuroscience, which began to link higher order cognition to its neural and even genetic substrate. In addition, new topics such as moral reasoning became active. After a few years, we decided the field of thinking and reasoning was ripe for a new comprehensive overview. This is it. Our aim is to provide comprehensive and authoritative reviews of all the core topics of the field of thinking and reasoning, with many pointers for further reading. Doubtless we still have omissions, but we have included as much as could realistically fit in a single volume. Our focus is on research from cognitive psychology, cognitive science, and cognitive neuroscience, but we also include work related to developmental, social and clinical psychology, philosophy, economics, artificial intelligence, linguistics, education, law, business,

and medicine. We hope that scholars and students in all these fields and others will find this to be a valuable collection.

We have many to thank for their help in bringing this endeavor to fruition. The editors at Oxford University Press, Catherine Carlin and more recently Joan Bossert, were instrumental in initiating and nurturing the project. We find it fitting that our new *Oxford Handbook of Thinking and Reasoning* should bear the imprint and indeed the name of this illustrious press, with its long history reaching back to the origins of scientific inquiry and its unparalleled list in the field of psychology. The entire staff at Oxford, especially Chad Zimmerman, provided us with close support throughout the arduous process of editing 40 chapters with 76 authors. During this period our own efforts were supported by grants from the Office of Naval Research (N000140810126), the Air Force Office of Scientific Research (FA9550-08-1-0489), and the Institute of Education Sciences (R305C080015) (KJH); and from the National Institute of Aging (T32AG020506), the Illinois Department of Public Health Alzheimer's Disease Research Fund, American Federation of Aging/Rosalinde and Arthur Gilbert Foundation, and the Loyola University Chicago Deans of Arts and Sciences and the Graduate School (RGM).

And then there are the authors. (It would seem a bit presumptuous to call them “our” authors!) People working on tough intellectual problems sometimes experience a moment of insight (see Chapter 24), a sense that although many laborious steps may lay ahead, the basic elements of a solution are already in place. Such fortunate people work on happily, confident that ultimate success is assured. In preparing this *Handbook*, we also had our moment of “insight.” It came when all these outstanding researchers had agreed to join our project. Before the first chapter was drafted, we knew the volume was going to be of the highest quality. Along the way, our distinguished authors graciously served as each other’s reviewers as we passed drafts around, nurturing each other’s chapters and adding in pointers from one to another. Then the authors all changed hats again and went back to work revising their own chapters in light of the feedback their peers had provided. We thank you all for making our own small labors a great pleasure.

Keith J. Holyoak
University of California, Los Angeles

Robert G. Morrison
Loyola University Chicago

CONTRIBUTORS

Paolo Ammirante

Department of Psychology
Ryerson University
Toronto, Canada

Jose F. Arocha

Department of Health Studies and
Gerontology
University of Waterloo
Waterloo, Ontario, Canada

Peter Bachman

Department of Psychiatry and
Biobehavioral Sciences
University of California, Los Angeles
Los Angeles, CA

Miriam Bassok

Department of Psychology
University of Washington
Seattle, WA

Mark Beeman

Department of Psychology
Northwestern University
Evanston, IL

Marc J. Buehner

School of Psychology
Cardiff University
Cardiff, Wales, UK

Colin F. Camerer

Division of Humanities and Social Sciences
Computation and Neural Systems
California Institute of Technology
Pasadena, CA

Tyrone D. Cannon

Departments of Psychology, Psychiatry and
Biobehavioral Sciences
University of California, Los Angeles
Los Angeles, CA

Alan D. Castel

Department of Psychology
University of California, Los Angeles
Los Angeles, CA

Nick Chater

Behavioural Science Group
Warwick Business School
University of Warwick
Coventry, UK

Patricia W. Cheng

Department of Psychology
University of California, Los Angeles
Los Angeles, CA

Susan Wagner Cook

Department of Psychology
University of Iowa
Iowa City, IA

Leonidas A. A. Doumas

Department of Psychology
University of Hawaii at Manoa
Manoa, HI

Kevin N. Dunbar

Department of Human Development
and Quantitative Methodology
University of Maryland
College Park, MD

Jonathan St. B. T. Evans

School of Psychology
University of Plymouth
Plymouth, UK

Jessica I. Fleck

Department of Psychology
The Richard Stockton College of
New Jersey
Galloway Township, NJ

Brandy N. Frazier

Department of Psychology
University of Hawaii at Manoa
Manoa, HI

Michael C. Friedman

Department of Psychology
University of California,
Los Angeles
Los Angeles, CA

Susan A. Gelman

Department of Psychology
University of Michigan
Ann Arbor, MI

Thomas Gilovich

Department of Psychology
Cornell University
Ithaca, NY

Lila Gleitman

Department of Psychology
University of Pennsylvania
Philadelphia, PA

Susan Goldin-Meadow

Department of Psychology
University of Chicago
Chicago, IL

Robert L. Goldstone

Department of Psychological and
Brain Sciences
Indiana University
Bloomington, IN

Richard Gonzalez

Department of Psychology
University of Michigan
Ann Arbor, MI

Adam E. Green

Department of Psychology and
Interdisciplinary Program in
Neuroscience
Georgetown University
Washington, DC

Dale W. Griffin

Sauder School of Business
University of British Columbia
Vancouver, British Columbia, Canada

Thomas L. Griffiths

Department of Psychology
University of California, Berkeley
Berkeley, CA

Ulrike Hahn

School of Psychology
Cardiff University
Cardiff, Wales, UK

Mary Hegarty

Department of Psychology
University of California,
Santa Barbara
Santa Barbara, CA

E. Tory Higgins

Departments of Psychology and
Management
Columbia University
New York, NY

Keith J. Holyoak

Department of Psychology
University of California,
Los Angeles
Los Angeles, CA

John E. Hummel

Department of Psychology
University of Illinois at
Urbana-Champaign
Champaign, IL

P. N. Johnson-Laird

Department of Psychology
Princeton University
Princeton, NJ

Charles Kemp

Department of Psychology
Carnegie Mellon University
Pittsburgh, PA

David Klahr

Department of Psychology
Carnegie Mellon University
Pittsburgh, PA

Barbara J. Knowlton

Department of Psychology
University of California, Los Angeles
Los Angeles, CA

Kenneth R. Koedinger

Departments of Human-Computer
Interaction and Psychology
Carnegie Mellon University
Pittsburgh, PA

Derek J. Koehler

Department of Psychology
University of Waterloo
Waterloo, Ontario, Canada

John Kounios

Department of Psychology
Drexel University
Philadelphia, PA

Robyn A. LeBoeuf

Marketing Department
University of Florida
Gainesville, FL

Jeffrey Loewenstein

Department of Business Administration
University of Illinois at
Urbana-Champaign
Champaign, IL

Tania Lombrozo

Department of Psychology
University of California, Berkeley
Berkeley, CA

Arthur B. Markman

Department of Psychology
University of Texas
Austin, TX

Shannon McGillivray

Department of Psychology
University of California,
Los Angeles
Los Angeles, CA

Douglas L. Medin

Department of Psychology
Northwestern University
Evanston, IL

Daniel C. Molden

Department of Psychology
Northwestern University
Evanston, IL

Robert G. Morrison

Department of Psychology
Loyola University Chicago
Chicago, IL

Jonas Nagel

Department of Psychology
University of Göttingen
Göttingen, Germany

Laura R. Novick

Department of Psychology and
Human Development
Vanderbilt University
Nashville, TN

Mike Oaksford

Department of Psychological
Sciences
Birkbeck College London
London, UK

John E. Opfer

Department of Psychology
The Ohio State University
Columbus, OH

Anna Papafragou

Department of Psychology
University of Delaware
Newark, DE

Vimla L. Patel

Center for Cognitive Studies
in Medicine and Public Health
New York Academy of Medicine
New York, NY

Derek C. Penn

Department of Psychology
University of California, Los Angeles
Los Angeles, CA

Daniel J. Povinelli

Cognitive Evolution Group
University of Louisiana
New Iberia, LA

Tage S. Rai

Department of Psychology
University of California, Los Angeles
Los Angeles, CA

Lance J. Rips

Department of Psychology
Northwestern University
Evanston, IL

Ido Roll

Carl Wieman Science Education
Initiative
Department of Physics and Astronomy
University of British Columbia
Vancouver, British Columbia, Canada

Frederick Schauer

School of Law
University of Virginia
Charlottesville, VA

Eldar Shafir

Department of Psychology and Woodrow
Wilson School of Public Affairs
Princeton University
Princeton, NJ

Robert S. Siegler

Department of Psychology
Carnegie Mellon University
Pittsburgh, PA

Dean Keith Simonton

Department of Psychology
University of California, Davis
Davis, CA

Alec Smith

Division of Humanities and Social Sciences
California Institute of Technology
Pasadena, CA

Edward E. Smith

Department of Psychology
Columbia University
New York, NY

Steven M. Smith

Department of Psychology
Texas A&M University
College Station, TX

Ji Yun Son

Department of Psychology
California State University, Los Angeles
Los Angeles, CA

Barbara A. Spellman

Department of Psychology and
School of Law
University of Virginia
Charlottesville, VA

Keith E. Stanovich

Department of Human Development and
Applied Psychology
University of Toronto
Toronto, Ontario, Canada

Andrew T. Stull

Department of Psychology
University of California, Santa Barbara
Santa Barbara, CA

Joshua B. Tenenbaum

Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Boston, MA

William Forde Thompson

Department of Psychology
Macquarie University
Sydney, Australia

J. Jason van Steenburgh

Department of Psychiatry
The Johns Hopkins School of Medicine
Baltimore, MD

Michael R. Waldmann

Department of Psychology
University of Göttingen
Göttingen, Germany

Thomas B. Ward

Department of Psychology
University of Alabama
Tuscaloosa, AL

Alex Wiegmann

Department of Psychology
University of Göttingen
Göttingen, Germany

Jiajie Zhang

Center for Cognitive Informatics
and Decision Making
School of Biomedical Informatics
University of Texas at Houston
Houston, TX

CONTENTS

1. Thinking and Reasoning: A Reader's Guide 1
Keith J. Holyoak and Robert G. Morrison

Part One • General Approaches to Thinking and Reasoning

2. Normative Systems: Logic, Probability, and Rational Choice 11
Nick Chater and Mike Oaksford
3. Bayesian Inference 22
Thomas L. Griffiths, Joshua B. Tenenbaum, and Charles Kemp
4. Knowledge Representation 36
Arthur B. Markman
5. Computational Models of Higher Cognition 52
Leonidas A. A. Doumas and John E. Hummel
6. Neurocognitive Methods in Higher Cognition 67
Robert G. Morrison and Barbara J. Knowlton
7. Mental Function as Genetic Expression: Emerging Insights From Cognitive Neurogenetics 90
Adam E. Green and Kevin N. Dunbar

Part Two • Deductive, Inductive, and Abductive Reasoning

8. Dual-Process Theories of Deductive Reasoning: Facts and Fallacies 115
Jonathan St. B. T. Evans
9. Inference in Mental Models 134
P. N. Johnson-Laird
10. Similarity 155
Robert L. Goldstone and Ji Yun Son
11. Concepts and Categories: Memory, Meaning, and Metaphysics 177
Lance J. Rips, Edward E. Smith, and Douglas L. Medin
12. Causal Learning 210
Patricia W. Cheng and Marc J. Buehner
13. Analogy and Relational Reasoning 234
Keith J. Holyoak
14. Explanation and Abductive Inference 260
Tania Lombrozo
15. Rational Argument 277
Ulrike Hahn and Mike Oaksford

Part Three • Judgment and Decision Making

16. Decision Making 301
Robyn A. LeBoeuf and Eldar Shafir
17. Judgmental Heuristics: A Historical Overview 322
Dale W. Griffin, Richard Gonzalez, Derek J. Koehler, and Thomas Gilovich
18. Cognitive Hierarchies and Emotions in Behavioral Game Theory 346
Colin F. Camerer and Alec Smith
19. Moral Judgment 364
Michael R. Waldmann, Jonas Nagel, and Alex Wiegmann
20. Motivated Thinking 390
Daniel C. Molden and E. Tory Higgins

Part Four • Problem Solving, Intelligence, and Creative Thinking

21. Problem Solving 413
Miriam Bassok and Laura R. Novick
22. On the Distinction Between Rationality and Intelligence: Implications for Understanding Individual Differences in Reasoning 433
Keith E. Stanovich
23. Cognition and the Creation of Ideas 456
Steve M. Smith and Thomas B. Ward
24. Insight 475
J. Jason van Steenburgh, Jessica I. Fleck, Mark Beeman, and John Kounios
25. Genius 492
Dean Keith Simonton

Part Five • Ontogeny, Phylogeny, Language, and Culture

26. Development of Thinking in Children 513
Susan A. Gelman and Brandy N. Frazier
27. The Human Enigma 529
Derek C. Penn and Daniel J. Povinelli
28. New Perspectives on Language and Thought 543
Lila Gleitman and Anna Papafragou
29. Thinking in Societies and Cultures 569
Tage S. Rai

Part Six • Modes of Thinking

30. Development of Quantitative Thinking 585
John E. Opfer and Robert S. Siegler
31. Visuospatial Thinking 606
Mary Hegarty and Andrew T. Stull
32. Gesture in Thought 631
Susan Goldin-Meadow and Susan Wagner Cook
33. Impact of Aging on Thinking 650
Shannon McGillivray, Michael C. Friedman, and Alan D. Castel

34. The Cognitive Neuroscience of Thought Disorder in Schizophrenia 673
Peter Bachman and Tyrone D. Cannon

Part Seven • Thinking in Practice

35. Scientific Thinking and Reasoning 701
Kevin N. Dunbar and David Klahr
36. Legal Reasoning 719
Barbara A. Spellman and Frederick Schauer
37. Medical Reasoning and Thinking 736
Vimla L. Patel, Jose F. Arocha, and Jiajie Zhang
38. Thinking in Business 755
Jeffrey Loewenstein
39. Musical Thought 774
William Forde Thompson and Paolo Ammirante
40. Learning to Think: Cognitive Mechanisms of Knowledge Transfer 789
Kenneth R. Koedinger and Ido Roll

Index 807

This page intentionally left blank

Thinking and Reasoning: A Reader's Guide

Keith J. Holyoak and Robert G. Morrison

“*Cogito, ergo sum*,” the French philosopher René Descartes famously declared, “I think, therefore I am.” Every fully functioning human adult shares a sense that the ability to think, to reason, is a part of one’s fundamental identity. A person may be struck blind or deaf yet still recognize his or her core cognitive capacities as intact. Even loss of language, the gift often claimed as the *sine qua non* of *Homo sapiens*, does not take away a person’s essential humanness. Perhaps thinking, not language, lies closest to both the core of our individual identity and to what is special about our species (see Penn & Povinelli, Chapter 27; Gleitman & Papafragou, Chapter 28). A person who loses language but can still make intelligent decisions, as demonstrated by actions, is viewed as mentally competent. In contrast, the kinds of brain damage that rob an individual of the capacity to think and reason are considered the harshest blows that can be struck against a sense of personhood (see Morrison & Knowlton, Chapter 6).

Cogito, ergo sum.

What Is Thinking?

We can start to answer this question by looking at the various ways the word *thinking* is used in everyday language. “I think that water is necessary for life” and “Keith and Bob think George was a fascist” both express *beliefs* (of varying degrees of apparent plausibility)—explicit claims of what someone takes to be a truth about the world. “Ann is sure to think of a solution” carries us into the realm of problem solving, the mental construction of an action plan to achieve a goal. The complaint, “Why didn’t you think before you went ahead with your half-baked scheme?” emphasizes that thinking can be a kind of *foresight*, a way of “seeing” the possible future.¹ “What do you think about it?” calls

for a *judgment*, an assessment of the desirability of an option. “Genocide is evil” takes judgment into the *moral* domain. And then there’s “Albert is lost in thought,” where thinking becomes some sort of mental meadow through which a person might meander on a rainy afternoon, oblivious to the world outside.

Rips and Conrad (1989) elicited judgments from college students about how various mentalistic terms relate to one another. Using statistical techniques, the investigators were able to summarize these relationships in two diagrams, shown in Figure 1.1. Figure 1.1A is a hierarchy of *kinds*, or categories. Roughly, people think planning is a kind of deciding, which is a kind of reasoning, which is a kind of conceptualizing, which is a kind of thinking. People also think (that verb again!) that thinking is *part of* conceptualizing, which is part of remembering, which is part of reasoning, and so on (Fig. 1.1B). The kinds ordering and the parts ordering are quite similar; most strikingly, *thinking* is the most general term in both orderings—the grand superordinate of mental activities, which permeates all the others.

Cogito, ergo sum.

It is not easy to make the move from the free flow of everyday speech to scientific definitions of mental terms, but let us nonetheless offer a preliminary definition of thinking to suggest what this book is about:

Thinking is the systematic transformation of mental representations of knowledge to characterize actual or possible states of the world, often in service of goals.

Obviously our definition introduces a plethora of terms with meanings that beg to be unpacked, but at which we can only hint. A *mental representation* of knowledge is an internal description that

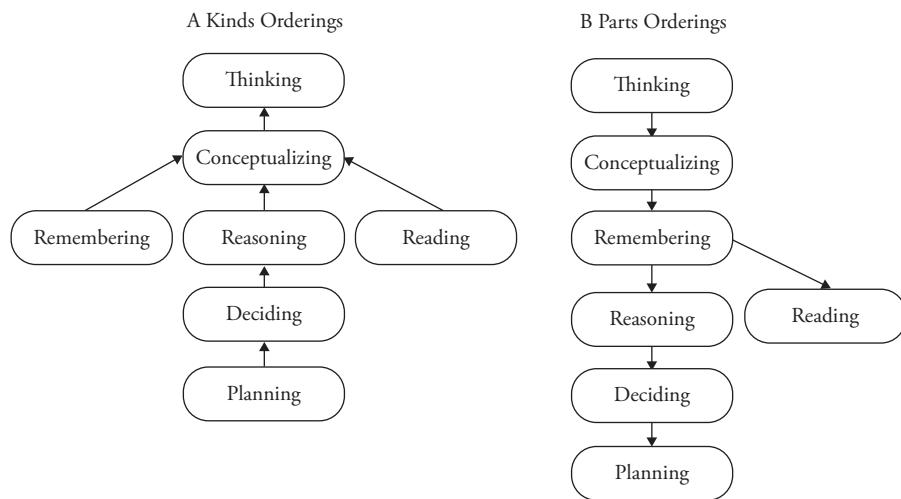


Fig. 1.1 People's conceptions of the relationships among terms for mental activities. (A) Ordering of "kinds." (B) Ordering of "parts." (Adapted from Rips & Conrad, 1989, with permission.)

can be manipulated to form other descriptions (see Markman, Chapter 4). To count as thinking, the manipulations must be *systematic* transformations that may be described computationally (see Doumas & Hummel, Chapter 5), governed by certain constraints. Whether a logical deduction (see Evans, Chapter 8) or a creative leap (see Smith & Ward, Chapter 23), what we mean by thinking is more than unconstrained associations (with the caveat that thinking may indeed be disordered; see Bachman & Cannon, Chapter 34). The internal representations created by thinking describe states of some external world (a world that may include the thinker as an object of self-reflection; see Gelman & Frazier, Chapter 26); that world might be our everyday one or perhaps some imaginary construction obeying the "laws" of magical realism. And often (not always—the daydreamer, and indeed the night dreamer, is also a thinker), thinking is directed toward achieving some desired state of affairs, some goal that motivates the thinker to do mental work (see Molden & Higgins, Chapter 20).

Our definition thus includes quite a few stipulations, but notice also what is left out. We do not claim that thinking necessarily requires a human or even a sentient being. Nonetheless, our focus in this book is on thinking by hominids with electrochemically powered brains constrained by their genes. Thinking often seems to be a conscious activity, of which the thinker is aware (*cogito, ergo sum*); but consciousness is a thorny philosophical puzzle, and some mental activities seem pretty

much like thinking except for being implicit rather than explicit (see Evans, Chapter 8). Finally, we do not claim that thinking is inherently rational, or optimal, or desirable, or even smart (see Stanovich, Chapter 22). A thorough history of human thinking will include quite a few chapters on stupidity; but at its pinnacle, thinking can be sheer genius (see Simonton, Chapter 25).

The study of thinking includes several interrelated subfields, which reflect slightly different perspectives on thinking. *Reasoning*, which has a long tradition that springs from philosophy and logic, places emphasis on the process of drawing inferences (*conclusions*) from some initial information (*premises*). In standard logic, an inference is *deductive* if the truth of the premises guarantees the truth of the conclusion by virtue of the argument form. If the truth of the premises renders the truth of the conclusion more credible, but does not bestow certainty, the inference is called *inductive*.² *Judgment and decision making* involve assessment of the value of an option or the probability that it will yield a certain payoff (judgment), coupled with choice among alternatives (decision making). *Problem solving* involves the construction of a course of action that can achieve a goal.

Although these distinct perspectives on thinking are useful in organizing the field (and this volume), these aspects of thinking overlap in every conceivable way. To solve a problem, one is likely to reason about the consequences of possible actions and to make decisions to select among alternative actions.

A logic problem, as the name implies, is a problem to be solved (with the goal of deriving or evaluating a possible conclusion). Making a decision is often a problem that requires reasoning. And so on. These subdivisions of the field, like our preliminary definition of thinking, should be treated as guideposts, not destinations.

A Capsule History

Thinking and reasoning, long the academic province of philosophy, have over the past century emerged as core topics of empirical investigation and theoretical analysis in the modern fields known as cognitive psychology, cognitive science, and cognitive neuroscience. Before psychology was founded, the 18th-century philosophers Immanuel Kant (in Germany) and David Hume (in Scotland) laid the foundations for all subsequent work on the origins of causal knowledge, perhaps the most central problem in the study of thinking (see Cheng & Buehner, Chapter 12). And if we were to choose one phrase to set the stage for modern views of thinking, it would be an observation of the British philosopher Thomas Hobbes, who in 1651 in his treatise *Leviathan* proposed “Reasoning is but reckoning.” *Reckoning* is an odd term today, but in the 17th century it meant “computation,” as in arithmetic calculations.³

It was not until the 20th century that the psychology of thinking became a scientific endeavor. The first half of the century gave rise to many important pioneers who in very different ways laid the foundations for the emergence of the modern field of thinking and reasoning. Foremost were the Gestalt psychologists of Germany, who provided deep insights into the nature of problem solving (see Bassok & Novick, Chapter 21; van Steenburgh et al., Chapter 24). Most notable of the Gestaltists were Karl Duncker and Max Wertheimer, students of human problem solving, and Wolfgang Köhler, a keen observer of problem solving by great apes.

The pioneers of the early 20th century also include Sigmund Freud, whose complex and ever-controversial legacy includes the notions that forms of thought can be unconscious, and that “cold” cognition is tangled up with “hot” emotion (see Molden & Higgins, Chapter 20). As the founder of clinical psychology, Freud’s legacy also includes the ongoing integration of research on “normal” thinking with studies of thought disorders, such as schizophrenia (see Bachman & Cannon, Chapter 34).

Other early pioneers in the early and mid-century contributed to various fields of study that are now

embraced within thinking and reasoning. Cognitive development (see Gelman & Frazier, Chapter 26) continues to be influenced by the early theories developed by the Swiss psychologist Jean Piaget and the Russian psychologist Lev Vygotsky. In the United States, Charles Spearman was a leader in the systematic study of individual differences in intelligence (see Stanovich, Chapter 22). In the middle of the century, the Russian neurologist Alexander Luria made immense contributions to our understanding of how thinking depends on specific areas of the brain, anticipating the modern field of cognitive neuroscience (see Morrison & Knowlton, Chapter 6). Around the same time in the United States, Herbert Simon argued that the traditional rational model of economic theory should be replaced with a framework that accounted for a variety of human resource constraints, such as bounded attention and memory capacity and limited time (see LeBoeuf & Shafir, Chapter 16). This was one of the contributions that in 1978 earned Simon the Nobel Prize in Economics.

In 1943, the British psychologist Kenneth Craik sketched the fundamental notion that a mental representation provides a kind of model of the world that can be “run” to make predictions (much like an engineer might use a physical scale model of a bridge to anticipate the effects of stress on the actual bridge intended to span a river).⁴ In the 1960s and 1970s, modern work on the psychology of reasoning began in Britain with the contributions of Peter Wason and his collaborator Philip Johnson-Laird (see Evans, Chapter 8; Johnson-Laird, Chapter 9).

The modern conception of thinking as computation became prominent in the 1970s. In their classic treatment of human problem solving, Allen Newell and Herbert Simon (1972) showed that the computational analysis of thinking (anticipated by Alan Turing, the father of computer science) could yield important empirical and theoretical results. Like a program running on a digital computer, a person thinking through a problem can be viewed as taking an input that represents initial conditions and a goal, and applying a sequence of operations to reduce the difference between the initial conditions and the goal. The work of Newell and Simon established computer simulation as a standard method for analyzing human thinking (see Doumas & Hummel, Chapter 5). It also highlighted the potential of production systems, which were subsequently developed extensively as cognitive models by John Anderson and his colleagues (see Koedinger & Roll, Chapter 40).

The 1970s saw a wide range of major developments that continue to shape the field. Eleanor Rosch, building on earlier work by Jerome Bruner (Bruner, Goodnow, & Austin, 1956), addressed the fundamental question of why people have the categories they do, and not other logically possible groupings of objects (see Rips, Smith, & Medin, Chapter 11). Rosch argued that natural categories often have fuzzy boundaries (a whale is an odd mammal), but nonetheless have clear central tendencies, or prototypes (people by and large agree that a bear makes a fine mammal). The psychology of human judgment was reshaped by the insights of Amos Tversky and Daniel Kahneman, who identified simple cognitive strategies, or heuristics, that people use to make judgments of frequency and probability. Often quick and accurate, these strategies can in some circumstances lead to nonnormative judgments. After Tversky's death in 1996, this line of work was continued by Kahneman, who was awarded the Nobel Prize in Economics in 2002. The current view of judgment that has emerged from 30 years of research is summarized by Griffin et al. (Chapter 17; also see LeBoeuf & Shafir, Chapter 16). Goldstone and Son (Chapter 10) review Tversky's influential theory of similarity judgments.

In 1982 David Marr, a young vision scientist, laid out a vision of how the science of mind should proceed. Marr distinguished three levels of analysis, which he termed the levels of *computation*, *representation and algorithm*, and *implementation*. Each level, according to Marr, addresses different questions, which he illustrated with the example of a physical device, the cash register. At Marr's most abstract level, computation (not to be confused with computation of an algorithm on a computer), the basic questions are "What is the goal that the cognitive process is meant to accomplish?" and "What is the logic of the mapping from the input to the output that distinguishes this mapping from other input-output mappings?" A cash register, viewed at this level, is used to achieve the goal of calculating how much is owed for a purchase. This task maps precisely onto the axioms of addition (e.g., the amount owed shouldn't vary with the order in which items are presented to the sales clerk, a constraint that precisely matches the commutativity property of addition). It follows that without knowing anything else about the workings of a particular cash register, we can be sure that (if it is working properly) it will be doing addition (not division).

The level of representation and algorithm, as the name implies, deals with the questions, "What is the representation of the input and output?" and "What is the algorithm for transforming the former into the latter?" Within a cash register, addition might be performed using numbers in either decimal or binary code, starting with either the leftmost or rightmost digit. Finally, the level of implementation addresses the question, "How are the representation and algorithm realized physically?" The cash register could be implemented as an electronic calculator, or a mechanical adding machine, or even a mental abacus in the mind of the clerk.

In his book, Marr stressed the importance of the computational level of analysis, arguing that it could be seriously misleading to focus prematurely on the more concrete levels of analysis for a cognitive task without understanding the goal or nature of the mental computation.⁵ Sadly, Marr died of leukemia before his book was published, so we do not know how his thinking about levels of analysis might have evolved. In very different ways, Marr's conception of a computational level of analysis is reflected in several chapters in this book (see especially Chater & Oaksford, Chapter 2; Griffiths, Tenenbaum, & Kemp, Chapter 3; Cheng & Buehner, Chapter 12; and Hahn & Oaksford, Chapter 15).

In the most recent quarter century many other springs of research have fed into the river of thinking and reasoning, including relational reasoning (see Holyoak, Chapter 13), neural network models (see Doumas & Hummel, Chapter 5), cognitive neuroscience (see Morrison & Knowlton, Chapter 6), and cognitive neurogenetics (Green & Dunbar, Chapter 7). The chapters of this *Handbook* collectively paint a picture of the state of the field in the early years of the new millennium.

Overview of the Handbook

This volume brings together the contributions of many of the leading researchers in thinking and reasoning to create the most comprehensive overview of research on thinking and reasoning that has ever been available. Each chapter includes a bit of historical perspective on the topic and ends with some thoughts about where the field seems to be heading. The book is organized into seven sections.

Part I: General Approaches to Thinking and Reasoning

The seven chapters in Part I address foundational issues. Chapter 2 by Chater and Oaksford lays out

the major normative theories (logic, probability, and rational choice) that have been used as standards against which human thinking and reasoning are often compared. In Chapter 3, Griffiths, Tenenbaum, and Kemp provide an overview of the Bayesian framework for probabilistic inference, which has been reinvigorated in recent years. Chapter 4 by Markman provides an overview of different conceptions of mental representation, and Chapter 5 by Doumas and Hummel surveys approaches to building computational models of thinking and reasoning. Then in Chapter 6, Morrison and Knowlton provide an introduction to the methods and findings of cognitive neuroscience as they bear on higher cognition, and in Chapter 7 Green and Dunbar discuss the emerging links between thinking and cognitive neurogenetics.

Part II: Inductive, Deductive, and Abductive Reasoning

Chapters 8–15 deal with core topics of reasoning. In Chapter 8, Evans reviews dual-process theories of reasoning, with emphasis on the psychology of deductive reasoning, the form of thinking with the closest ties to logic. In Chapter 9, Johnson-Laird describes the work that he and others have done using the framework of mental models to deal with various reasoning tasks, both deductive and inductive. Chapter 10 by Goldstone and Son reviews work on the core concept of similarity—how people assess the degree to which objects or events are alike. Chapter 11 by Rips, Smith, and Medin considers research on categories, and how concepts are organized in semantic memory. In Chapter 12, Cheng and Buehner discuss causal learning, a basic type of induction concerning how humans and other creatures acquire knowledge about causal relations, which are critical for predicting the consequences of actions and events. Then, in Chapter 13, Holyoak reviews the literature on reasoning by analogy and similar forms of relational reasoning. In Chapter 14, Lombrozo explores the multifaceted topic of explanation, which is closely related to abductive reasoning (often called “inference to the best explanation”). Then, in Chapter 15, Hahn and Oaksford apply the Bayesian framework to understand how people interpret informal arguments, including types of arguments that have classically been viewed as logical fallacies.

Part III: Judgment and Decision Making

In Chapters 16–20 we turn to topics related to judgment and decision making. In Chapter 16,

LeBoeuf and Shafir set the stage with a general review of work on decision making. Then, in Chapter 17, Griffin, Gonzalez, Koehler and Gilovich review the fascinating literature on heuristics and biases that influence judgment. In Chapter 18, Camerer and Smith discuss behavioral game theory, an approach rooted in economics that has been applied in many other disciplines. They also touch upon recent work on neuroeconomics, the study of the neural substrate of decision making. In Chapter 19, Waldmann, Nagel, and Wiegmann review a growing literature on moral reasoning and decision making. Then, in Chapter 20, Molden and Higgins review research revealing the ways in which human motivation and emotion influence judgment.

Part IV: Problem Solving, Intelligence, and Creative Thinking

The five chapters that comprise this section deal with problem solving and the many forms of individual differences observed in human thinking. In Chapter 21, Bassok and Novick provide a general overview of the field of human problem solving. In Chapter 22, Stanovich analyzes different conceptions of rationality and discusses individual differences in both rational thought and intelligence. Problem solving has close connections to the topic of creativity, the focus of Chapter 23 by Smith and Ward. In Chapter 24, van Steenburgh, Fleck, Beeman, and Kounios review research that takes a cognitive neuroscience approach to understanding the basis for insight in problem solving. Finally, in Chapter 25 Simonton reviews what is known about the thinking processes of those who function at the extreme of individual differences commonly termed “genius.”

Part V: Ontogeny, Phylogeny, Language, and Culture

Our understanding of thinking and reasoning would be gravely limited if we restricted investigation to young adult English speakers. Chapters 26–29 deal with the multifaceted ways in which aspects of thinking vary across the human life span, across species, across speakers of different languages, and its connections to larger human groups. In Chapter 26, Gelman and Frazier provide an overview of the development of thinking and reasoning over the course of childhood. In Chapter 27, Penn and Povinelli consider the fundamental question of what makes human thinking special when compared to the mental functioning of nonhuman animals. One of the most controversial topics in the field is the relationship

between thinking and the language spoken by the thinker. In Chapter 28, Gleitman and Papafragou offer a fresh perspective on the hypotheses and evidence concerning the connections between language and thought. Finally, in Chapter 29, Rai discusses the ways in which human thinking can be viewed as distributed across social and cultural groups.

Part VI: Modes of Thinking

There are many modes of thinking, distinguished by broad variations in representations and processes. Chapters 30–34 consider a number of these. In Chapter 30, Opfer and Siegler discuss mathematical thinking, a special form of thinking found in rudimentary form in nonhuman animals and which undergoes complex developmental changes over the course of childhood. In Chapter 31, Hegarty and Stull review work on the role of visuospatial representations in thinking; and in Chapter 32, Goldin-Meadow and Cook discuss the ways in which spontaneous gestures reflect and guide thinking processes. In Chapter 33, McGillivray, Friedman, and Castel describe the changes in thinking and reasoning brought on by the aging process. In Chapter 34, Bachman and Cannon review research and theory concerning brain disorders, notably schizophrenia, that produce striking disruptions of normal thought processes.

Part VII: Thinking in Practice

In cultures ancient and modern, thinking is put to particular uses in special cultural practices. Chapters 35–40 focus on thinking in particular practices. In Chapter 35, Dunbar and Klahr discuss thinking and reasoning as manifested in the practice of science. In Chapter 36, Spellman and Schauer review different conceptions of legal reasoning. In Chapter 37, Patel, Arocha, and Zhang discuss reasoning in a field—medicine—in which accurate diagnosis and treatment is literally an everyday matter of life and death. Lowenstein discusses reasoning as it relates to business in Chapter 38. Thinking is also involved in many aspects of music, including composition; this topic is covered by Thompson and Ammirante in Chapter 39. Finally, Chapter 40 by Koedinger and Roll concludes the volume by considering one of the major challenges for education—finding ways to teach people to think more effectively.

Examples of Chapter Assignments for a Variety of Courses

The present volume offers a comprehensive treatment of higher cognition. As such, it serves as an

excellent source for courses on thinking and reasoning, both at the graduate level and for advanced undergraduates. While instructors for semester-length graduate courses in thinking and reasoning may opt to assign the entire volume as a textbook, there are a number of other possibilities (including using chapters from this volume as introductions for various topics and then supplementing with readings from the primary literature). Here are a few examples of possible chapter groupings, tailored to a variety of possible course offerings.

Introduction to Thinking and Reasoning

1. Thinking and Reasoning: A Reader's Guide
2. Normative Systems: Logic, Probability, and Rational Choice
3. Bayesian Inference
4. Knowledge Representation
8. Dual-Process Theories of Reasoning: Facts and Fallacies
9. Inference in Mental Models
10. Similarity
11. Concepts and Categories: Memory, Meaning, and Metaphysics
12. Causal Learning and Inference
13. Analogy and Relational Reasoning
14. Explanation and Abductive Inference
15. Rational Argument
16. Decision Making
17. Judgment Heuristics
21. Problem Solving
22. On the Distinction Between Rationality and Intelligence: Implications for Understanding Individual Differences in Reasoning
23. Cognition and the Creation of Ideas

Development of Thinking

1. Thinking and Reasoning: A Reader's Guide
4. Knowledge Representation
10. Similarity
11. Concepts and Categories: Memory, Meaning, and Metaphysics
13. Analogy and Relational Reasoning
14. Explanation and Abductive Inference
26. Development of Thinking in Children
27. The Human Enigma
28. Language and Thought
30. Mathematical Cognition
31. Visuospatial Thinking
32. Gesture in Thought

33. Impact of Aging on Thinking
40. Learning to Think: Cognitive Mechanisms of Knowledge Transfer

Modeling Human Thought

1. Thinking and Reasoning: A Reader's Guide
3. Bayesian Inference
4. Knowledge Representation
5. Computational Modeling of Higher Cognition
6. Neural Substrate of Thinking
9. Inference in Mental Models
10. Similarity
11. Concepts and Categories: Memory, Meaning, and Metaphysics
12. Causal Learning and Inference
13. Analogy and Relational Reasoning
15. Rational Argument
18. Cognitive Hierarchies and Emotions in Behavioral Game Theory
40. Learning to Think: Cognitive Methods of Knowledge Transfer

Applied Thought

1. Thinking and Reasoning: A Reader's Guide
35. Scientific Thinking and Reasoning
36. Legal Reasoning
37. Thinking and Reasoning in Medicine
38. Thinking in Business
39. Musical Thought
40. Learning to Think: Cognitive Methods of Knowledge Transfer

Differences in Thought

1. Thinking and Reasoning: A Reader's Guide
19. Moral Judgment
20. Motivated Thinking
23. Cognition and the Creation of Ideas
24. Insight
25. Genius
26. Development of Thinking in Children

27. The Human Enigma
28. Language and Thought
29. Thinking in Society and Culture
32. Gesture in Thought
33. Impact of Aging on Thinking
34. The Cognitive Neuroscience of Thought Disorder in Schizophrenia

Acknowledgments

Preparation of this chapter was supported by grants from the Office of Naval Research (N000140810186) and the Institute of Education Sciences (R305C080015) (KJH); and by the National Institute of Aging (T32AG020506), the Illinois Department of Public Health Alzheimer's Disease Research Fund, and the Loyola University Chicago Deans of Arts and Sciences and the Graduate School (RGM).

Notes

1. Notice the linguistic connection between "thinking" and "seeing," thought and perception, which was emphasized by the Gestalt psychologists of the early 20th century.
2. The distinction between deduction and induction blurs in the study of the psychology of thinking, as we will see in Part II of this volume.
3. There are echoes of the old meaning of *reckon* in such phrases as "reckon the cost." As a further aside, the term "dead reckoning," a procedure for calculating the position of a ship or aircraft, derives from "deductive reasoning." And in an old Western movie, a hero in a tough spot might venture, "I reckon we can hold out till sun-up," illustrating how calculation has crossed over to become a metaphor for mental judgment.
4. See Johnson-Laird, Chapter 9, for a current view of thinking and reasoning that owes much to Craik's seminal ideas.
5. Indeed, Marr criticized Newell and Simon's approach to problem solving for paying insufficient attention to the computational level in this sense.

References

- Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.
- Craik, K. (1943) *The nature of explanation*. Cambridge, England: Cambridge University Press.
- Hobbes, T. (1651/1968). *Leviathan*. London: Penguin Books.
- Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
- Rips, L. J., & Conrad, F. G. (1989). Folk psychology of mental activities. *Psychological Review*, 96, 187–207.

This page intentionally left blank

PART

1

General Approaches
to Thinking and
Reasoning

This page intentionally left blank

Normative Systems: Logic, Probability, and Rational Choice

Nick Chater and Mike Oaksford

Abstract

Normative theories of how people *should* reason have been central to the development of the cognitive science of thinking and reasoning, both as standards against which how thought is assessed and as sources of hypotheses about how thought might operate. This chapter sketches three particularly important types of normative system: logic, probability, and rational choice theory, stressing that these can each be viewed as providing consistency conditions on thought. From the perspective of understanding thought, logic can be viewed as providing consistency conditions on *beliefs*; probability provides consistency conditions on *degrees of beliefs*; and rational choice provides consistency conditions on *choices*. Limitations of current normative approaches are discussed. Throughout this chapter, we provide pointers concerning how these ideas link with the empirical study of thinking and reasoning, as described in this book.

Key Words: logic, probability, rational choice, decision theory, consistency, normative explanation

Introduction

This volume addresses one of the central questions in psychology: how people think and reason. Like any scientific endeavor, the psychology of thinking and reasoning is concerned not with how things should be, but with how things actually are. So the psychological project is not directly concerned with determining how people *ought* to reason; all that matters is figuring out how they *do* reason.

This chapter is, nonetheless, concerned purely with *normative* questions, that is, questions of how reasoning *should* be conducted—it focuses on normative theories, including logic, probability theory, and rational choice theory. These normative questions are traditionally discussed in philosophy and mathematics, rather than science. But questions about how people *should* reason are important for creating a cognitive science of how people *do* reason, in at least three ways.

First of all, a normative theory can provide a standard against which actual behavior can be assessed. That is, it provides an analysis of what the system is “supposed” to do: It determines what counts as successful performance and what counts as error. Note, though, that it may not be straightforward to determine which normative theory is appropriate, in relation to a particular aspect of thought or behavior; or how that normative theory should be applied. Nonetheless, comparison with normative theories is crucial for framing many of the key questions concerning thinking and reasoning. For example, even to ask whether, or to what extent, people reason deductively (see Evans, Chapter 8) requires an analysis of what deduction is; and deduction is a *logical* notion. Similarly, the study of decision making (see LeBoeuf & Shafir, Chapter 16) is organized around comparisons with the theory of “correct” decision making, deriving from rational choice theory.

Second, normative theories of reasoning provide a possible starting point for descriptive theories of reasoning, rather than a mere standard of comparison. Thus, just as a natural starting point for a theory of the operation of a calculator is that it follows, perhaps to some approximation, the laws of arithmetic, it is natural to assume that a rational agent follows, to some approximation, normative principles of reasoning. This second use of normative theories is by far the most controversial. Many theories in the psychology of reasoning are built on foundations from normative theories. For example, mental logics and mental models view the reasoner as approximating logical inference (Johnson-Laird, 1983; Rips, 1994; see Evans, Chapter 8; Johnson-Laird, Chapter 9) and Bayesian approaches view cognition as approximating Bayesian inference (Griffiths, Kemp, & Tenenbaum, 2008; Oaksford & Chater, 2007; see Griffiths, Tenenbaum, & Kemp, Chapter 3). But other accounts, based, for example, on heuristics (e.g., Evans, 1984; Gigerenzer & Todd, 1999), take this to be a fundamental mistake.

Third, unless people's words and behavior obey some kind of rationality constraints, their words and behavior are rendered, quite literally, meaningless. Thus, a person randomly generating sentences, or moving arbitrarily, cannot usefully be attributed intentions, beliefs, or goals, at least according to many influential philosophical accounts (e.g., Davidson, 1984; Quine, 1960). To put the point starkly: to see a human body as a *person*, with a *mind*, rather than merely a collection of physiological responses, requires that he or she can be attributed some measure of rationality. And normative theories of rationality attempt to make some headway in explaining the nature of such rationality constraints. Similarly, returning to comparison with the calculator, without some degree of conformity with the laws of arithmetic, it will be impossible to interpret the key presses and displays of the calculator as being about numbers or arithmetic operations.

A natural initial question is: Is it meaningful to attempt to develop a general theory of rationality *at all*? We might tentatively suggest that it is a *prima facie* sign of irrationality to believe in alien abduction, or to will a sports team to win in order to increase their chance of victory. But these views or actions might be entirely rational, given suitably nonstandard background beliefs about other alien activity and the general efficacy of psychic powers. Irrationality may, though, be ascribed if there is a *clash* between a particular belief or behavior and such

background assumptions. Thus, a thorough-going physicalist may, perhaps, be accused of irrationality if she simultaneously believes in psychic powers. A theory of rationality cannot, therefore, be viewed as clarifying either what people should believe or how people should act—but it can determine whether beliefs and behaviors are compatible. Similarly, a theory of rational choice cannot determine whether it is rational to smoke or to exercise daily; but it might clarify whether a particularly choice is compatible with other beliefs and choices.

From this viewpoint, normative theories can be viewed as clarifying conditions of *consistency* (this is a broader notion than “logical consistency,” which we shall introduce later). In this chapter, we will discuss theories that aim to clarify notions of consistency in three domains. *Logic* can be viewed as studying the notion of consistency over *beliefs*. *Probability* (from the subjective viewpoint we describe shortly) studies consistency over *degrees of belief*. *Rational choice* theory studies the consistency of beliefs and values with *choices*. In each domain, we shall interpret the formalisms in a way most directly relevant to thought. Logic can be applied to formalizing mathematical proofs or analyzing computer algorithms; probability can capture limiting frequencies of repeatable events; rational choice theory can be applied to optimization problems in engineering. But we shall focus, instead, on the application of these formal theories to providing models of thought.

Logic

Our intuitions about the rationality of any single belief crucially may be influenced by background beliefs. This type of *global* relationship, between a single belief and the morass of general background knowledge, will inevitably be hard to analyze. But logic focuses on *local* consistency (and consequence) relationships between beliefs, which depend on those beliefs alone, so that background knowledge is irrelevant. Such local relationships can only depend on the *structure* of the beliefs (we'll see what this means shortly); it can't depend on what the beliefs are about, because understanding what a belief is *about* requires reference to background knowledge (and perhaps other things besides).

So, for example, there seems something wrong, if Fred believes the following:

All worms warble
Albert is a worm
Albert does not warble

Roughly, the first two beliefs appear to imply that Albert does warble; and yet the third belief is that he does not. And surely a minimal consistency condition for any reasoner is that it should not be possible to believe both P & not P —because this is a *contradiction*. This inconsistency holds completely independently of any facts or beliefs about worms, warbling, or Albert—this is what makes the inconsistency *local*.

How can these intuitions be made more precise? Logic plays this role, by translating sentences of natural language (expressing beliefs) into formulae of a precisely defined formal language; and specifying inferential rules over those formulae. Sometimes there are also axioms as well as rules; we will use rules exclusively here for simplicity, thus following the framework of natural deduction in logic (Barwise & Etchemendy, 2000). The idea that human reasoning works roughly in this way is embodied in mental logic approaches in the psychology of reasoning (e.g., Braine, 1978; Rips, 1994).

Beliefs can be represented in many different logical languages; and many different sets of inferences can be defined over these languages. So there is not one logic, but many. Let us represent the beliefs mentioned earlier in perhaps the best known logical language, the first-order predicate calculus:

All worms warble	$\forall x. (\text{worm}(x) \rightarrow \text{warble}(x))$
Albert is a worm	$\text{worm}(\text{Albert})$
Albert does not warble	$\neg\text{warble}(\text{Albert})$

where “ $\forall x$ ” can be glossed as “for all x ,” and “ \rightarrow ” can be glossed as “if...then...” (although with a very particular interpretation; different interpretations of the conditional are numerous and contested, e.g., Edgington, 1995; Evans & Over, 2004), and “ \neg ” can be glossed as “not” (i.e., as negating the following material within its scope). The variable x ranges over a set of objects; and Albert is one of these objects.

Now there is a logical rule (\forall -elimination) which holds in the predicate calculus; this captures the intuition that, if some formulae applies to any object (any x), then it must apply to any particular object, such as Albert. Applying this rule to $\forall x. (\text{worm}(x) \rightarrow \text{warble}(x))$, we obtain:

$$\text{worm}(\text{Albert}) \rightarrow \text{warble}(\text{Albert})$$

Then we can apply a second logical rule, \rightarrow -elimination, which says (simplifying slightly), that, for any beliefs P, Q , if it is true that $P \rightarrow Q$ and it

is true that P , then Q is also true. Since we already know $\text{worm}(\text{Albert})$, and that $\text{worm}(\text{Albert}) \rightarrow \text{warble}(\text{Albert})$, this rule tells us that we can derive $\text{warble}(\text{Albert})$.

So, with this reasoning, our first two premises have derived $\text{warble}(\text{Albert})$; and the third premise is $\neg\text{warble}(\text{Albert})$. To make the contradiction explicit, we need to apply a final logical rule, $\&$ -introduction, which states that, if we have established P , and Q , we can derive $P \& Q$. Thus, we can derive $\text{warble}(\text{Albert}) \& \neg\text{warble}(\text{Albert})$, which is, of course, of the form $P \& \neg P$, that is, a contradiction.

Thus, from the point of view of the cognitive science of reasoning, logic can be viewed as providing a formal mechanism for detecting such inconsistencies: by determining the sets of beliefs from which a contradiction can be derived. (The field of logic does not exhaust such methods. For example, the denial of Fermat’s Last Theorem may be in contradiction with the rules of arithmetic, but the detection of such an inconsistency, by proving Fermat’s Last Theorem, requires mathematics far outside the confines of logic.) The chains of reasoning that result from applying logical rules step by step is studied by one major branch of mathematical logic: proof theory.

But what makes the logical rules (\forall -elimination, $\&$ -introduction, and so on) appropriate? One answer is that, ideally, these rules should allow the derivation $P \& \neg P$ from a set of logical formulae just when that set of formulae is incompatible from the point of view of *meaning*. That is, whatever the properties of *warbling* or being a *worm* refer to, and whoever or whatever is denoted by *Albert*, our three statements cannot be true together. It turns out that, for the purposes of figuring out whether it is possible for a set of beliefs to be simultaneously true or not, under some interpretation, it is often sufficient to restrict our interpretations to sets of numbers (so we don’t have to worry about worms, warbling, or any other real or imagined features of the actual world), though more complex mathematical structures may be appropriate for other logics.

From this standpoint, our formulae

$$\begin{aligned} &\forall x. (\text{worm}(x) \rightarrow \text{warble}(x)) \\ &\text{worm}(\text{Albert}) \\ &\neg\text{warble}(\text{Albert}) \end{aligned}$$

can be true simultaneously just when there is some (potentially infinite) set of numbers S_{worm} denoted by *worm* and another S_{warble} by *warble*; and a number a denoted by *Albert*, such that each of these formulae

are true together. This can be done by defining a precise relationship between formulae and numbers and sets. Roughly, the first formula translates into $S_{\text{worm}} \subset S_{\text{warble}}$, that is, the first set of numbers is a subset of the second, following the possibility that the sets might be identical. The second translates as $a \in S_{\text{worm}}$, and the third as $a \notin S_{\text{warble}}$. There are clearly no numbers and sets of numbers that can make this true: $S_{\text{worm}} \subset S_{\text{warble}}$ requires that every number that is in S_{worm} is also in S_{warble} , and a provide a counterexample. We say that this set of formulae has no *model* in set-theory, that is, no set-theoretic interpretation that makes all of the formulae true. Conversely, if the first formula were, instead, the innocuous $\text{warble}(Albert)$, then the formulae would have a model (indeed, infinitely many such models). For example, Albert could be the number 5; S_{worm} could be the set {3, 5, 8}; and S_{warble} the set {3, 4, 5, 8}. If a set of formulae has a model (i.e., an interpretation under which all the statements are true), then it is *satisfiable*; otherwise it is *unsatisfiable*. The study of the relationship between formulae and possible interpretations in terms of sets (or in terms of more complex structures such as fields and rings) is the subject of a second major branch of mathematical logic: model theory.

So we now have introduced two notions of consistency, one proof-theoretic (using the logical rules, can we derive a contradiction of the form $P \& \neg P$?), and the other model-theoretic (are the formulae satisfiable, i.e., is there some model according to which they can all be interpreted as true?). Ideally, these notions of consistency should come to the same thing.

We say that a logical system is *sound* (with respect to some interpretation) just when its inference rules can derive a contradiction of the form $P \& \neg P$ from a set of formulae only when that set of formulae is unsatisfiable. Informally, soundness implies that proof theory respects model theory. Even more informally: If the logical rules allow you to derive a contradiction from some set of formulae, then they really can't all be true simultaneously.

Conversely, a logical system is *complete* (with respect to some interpretation) just when a set of formulae is unsatisfiable only when the inference rules can derive a contradiction of the form $P \& \neg P$ from that set of formulae. Informally, completeness implies that the interpretation in the model theory is captured by the proof theory. Even more informally: If a set of formulae can't all be true simultaneously, then the logical rules allow a contradiction to be derived from them.

We have introduced logic so far as a calculus of *consistency*, both to clarify the comparison with probability and rational choice theory and to illustrate some key points concerning the relationship between logic and reasoning, which we will discuss shortly. But the more conventional starting point, particularly appropriate given the roots of mathematical logic in clarifying the foundations of mathematics, is that logic provides an account of *consequence*. According to this standpoint, we consider a set of premises

$$\begin{aligned} &\forall x. (\text{worm}(x) \rightarrow \text{warble}(x)) \\ &\text{worm}(Albert) \end{aligned}$$

and wonder whether the conclusion $\text{warble}(Albert)$ follows. The link with consistency is direct: $\text{warble}(Albert)$ follows from these premises if and only if adding $\neg \text{warble}(Albert)$ to those premises leads to a contradiction. More generally, any argument from a set of premises Γ to the consequence P is *syntactically valid* (i.e., P can be derived from Γ , by applying the rules of inference) just when $\Gamma, \neg P$ allows a contradiction to be derived. And on the model-theoretic side, there is a corresponding notion of consequence: An argument from Γ to P is *semantically valid* just when any models satisfying Γ also satisfy Γ, P . And this will be true just when no model satisfies $\Gamma, \neg P$ —that is, $\Gamma, \neg P$ is unsatisfiable. So we can explain soundness and completeness in a different, although equivalent, way: A logic is sound when syntactically valid conclusions are always also semantically valid; and complete when the converse holds. But we shall continue primarily to think of terms of consistency in the following discussion.

Happily, standard predicate calculus (a fragment of which we used earlier) is both sound and complete (with respect to a standard set-theoretic interpretation), as Gödel showed (Boolos & Jeffrey, 1980). More powerful logics (e.g., those that allow the representation of sufficiently large fragments of arithmetic) can be shown not to be both sound and complete (Gödel's incompleteness theorem is a key result of this kind; Nagel & Newman, 1958). A logic that is not sound is not very useful—the proof theory will generate all kinds of “theorems” that need not actually hold. Sound but incomplete logics can, though, be very useful. For example, in mathematics, they can be used to prove interesting theorems, even if there are theorems that they cannot prove.

Let us stress again that (in)consistency is, crucially, a *local* notion. That is, a set of beliefs is

(in)consistent (when represented in some logical system) dependent only on those beliefs—and completely independent of background knowledge. This is entirely natural in mathematics—where we are concerned with what can be derived from our given axioms and crucially want to see what follows from these axioms alone. But where reasoning is not about mathematics, but about the external world, our inferences may be influenced by background knowledge—for example, in understanding a news story or a conversation, we will typically need to draw on large amounts of information not explicitly stated in the words we encounter. Our conclusions will not follow with certainty from the stated “premises,” but will, rather, be tentative conjectures based on an attempt to integrate new information with our background knowledge. Such reasoning does not seem to fit easily into a logical framework (Oaksford & Chater, 2007).

Moreover, consistency is also a *static* notion. Suppose we imagine that we believe the three previous statements—but that, on reflection, we realize that these are inconsistent. So at least one of these beliefs must be jettisoned. But which? Standard logic says nothing about this (Harman, 1986; see also, Macnamara, 1986). The *dynamic* question, of how we should modify our beliefs in order to restore consistency, is much studied—but it is a fundamentally different question from whether there is inconsistency in the first place.

As we have stressed, for this reason, logic cannot be viewed as a theory of reasoning: that is, as a theory of how beliefs should be changed, in the light of reflection. Suppose I believe that *worms warble* and that *Albert is a worm*. The conclusion that *Albert warbles* follows from these beliefs. But suppose I initially doubt that Albert warbles. Then, if I decide to hold to the first two beliefs, then I had better conclude that Albert does warble after all. But I might equally well maintain my doubt that Albert warbles and begin to suspect that not all worms warble after all; or perhaps Albert is not actually a worm. The key point is that logic tells us when inconsistency has struck, but it does not tell us how it can be restored. In mathematics, for which logic was developed, it is usual to take certain axioms (e.g., of geometry, set theory, or group theory) for granted—and then to accept whatever can be derived from them. But outside mathematics, no belief has protected status—and hence the question of how to restore consistency is a challenging one.

We have seen that logic has, on the face of it, an important, but limited, role in a theory of reasoning: focusing on local consistency relations. Nonetheless, various theorists have considered logic to play a central psychological role. Inhelder and Piaget (1955) viewed cognitive development as the gradual construction of richer logical representations; symbolic artificial intelligence modeled thought in terms of variants of logical systems, including frames, scripts, and schemas (Minsky, 1977; Schank & Abelson, 1977), and logical systems are still the basis for many theories of mental representation (see Chapter 4). Moreover, Fodor and Pylyshyn (1988) have argued that the central hypothesis of cognitive science is to view “cognition as proof theory,” a viewpoint embodied in the psychological proposals of “mental logic” (Braine, 1978; Rips, 1994).

An alternative proposal takes model-theoretic entailment as its starting point, proposing that people reason by formulating “mental models” of the situations being described (Johnson-Laird, 1983; Chapter 9, this volume; Johnson-Laird & Byrne, 1991). Possible conclusions are “read off” these mental models; and there is a search for alternative mental models that may provide counter-examples to these conclusions.

We have focused here on predicate logic, dealing with the analysis of terms such as *not*, *or*, *and*, *if...then...*, *all*, and *some*. But the variety of logics is very great: There are deontic logics, for reasoning about moral permissibility and obligation; modal logics for reasoning about possibility and necessity; temporal logics for modeling tense and aspect; second-order logics for reasoning about properties; and so on. Each such logic aims to capture a complementary aspect of the structure of beliefs; and one might hope that a fusion of these logics might capture these different features simultaneously. Typically, though, such fusion is very difficult—different aspects of logical structure are usually studied separately (Montague Grammar provides a partial exception; see Dowty, Wall, & Peters, 1981). Moreover, with regard to any natural language term, there are typically a variety of possible logical systems that may be applied, which do not yield the same results. The psychological exploration of the variety of logical systems is relatively underdeveloped, although the mental models approach has been applied particularly broadly (e.g., Johnson-Laird, Chapter 9). And, more generally, the degree to which people are sensitive at all to the kinds of local consistency relations described by logic, or indeed other normative theories, is a matter of controversy

(e.g., Gigerenzer & Todd, 1999; Oaksford & Chater, 2007; Rips, 1994).

Probability

Logic can be viewed as determining which sets of *beliefs* are (in)consistent. But belief may not be an all-or-none matter. Human thinking is typically faced with uncertainty. The senses provide information about the external world; but this information is not entirely dependable. Reports, even from a trustworthy informant, cannot be entirely relied upon. So, outside mathematics, we might expect belief to be a matter of degree.

Probability theory can be viewed as capturing consistency of *degrees of belief*. But what exactly is a degree of belief? It turns out that surprisingly few assumptions uniquely fix a calculus for degrees of belief—which are just the laws of probability. And different assumptions lead to the same laws of probability (e.g., Cox, 1946; Kolmogorov, 1956; Ramsey, 1926).

Viewing probability as modeling degrees of belief is, in philosophy, known as the *subjective* interpretation of probability. In statistics, machine learning, and, by extension, the cognitive sciences, this is typically known as the *Bayesian* approach (see Griffiths, Tenenbaum, & Kemp Chapter 3; Oaksford & Chater, 2007).

The Bayesian approach is named after Bayes theorem, because of the frequent appeal to the theorem in uncertain inference. While the Bayesian approach to probability and statistics is controversial (probabilities may, for example, be interpreted as limiting frequencies; see von Mises, 1939), and statistics may be viewed as involving the sequential rejection of hypotheses (see Fisher, 1925), Bayes theorem is not: It is an elementary theorem of probability.

In its very most basic form, Bayes theorem arises directly from the very definition of conditional probability: $\text{Pr}(A|B)$ is the probability that A is true, given that B is true. Then, it follows ineluctably that the joint probability $\text{Pr}(A, B)$, that both A and B are true, is simply the probability that B is true ($\text{Pr}(B)$) multiplied by the probability that A is true, given that B is true, $\text{Pr}(A|B)$: that is, $\text{Pr}(A, B) = \text{Pr}(A|B)\text{Pr}(B)$. By the symmetry roles of A and B , it also follows that $\text{Pr}(A, B) = \text{Pr}(B|A)\text{Pr}(A)$, and so that $\text{Pr}(B|A)\text{Pr}(A) = \text{Pr}(A|B)\text{Pr}(B)$. Dividing through by $\text{Pr}(A)$ gives a simple form of Bayes theorem:

$$\text{Pr}(B|A) = \frac{\text{Pr}(A|B)\text{Pr}(B)}{\text{Pr}(A)}$$

Bayes theorem plays a crucial role in probabilistic approaches to cognition. Often, $\text{Pr}(A|B)$ is known (e.g., the probability of some pattern data arising, if some hypothesis is true); but the “converse” $\text{Pr}(B|A)$ (e.g., the probability that the hypothesis is true, given that we are observed the pattern of data) is not known. Simplifying somewhat, Bayes theorem helps derive the unknown probability from known probabilities.

Given the axioms of probability (and the ability to derive corollaries of these axioms, such as Bayes theorem), the whole of probability theory is determined—and, at the same time, one might suppose, the implications of probability for cognitive science. But, as in other areas of mathematics, merely knowing the axioms is really only the starting point: the consequences of the axioms turn out to be enormously rich. Of particular interest to cognitive science is the fact that it turns out to be very useful to describe some cognitively relevant probability distributions in terms of graphical models (see Pearl, 1988), which are both representationally and computationally attractive. These models are beyond the scope of this chapter (but see Griffiths, Tenenbaum, & Kemp, Chapter 3). Note though that much current work on the representation of knowledge (see Markman, Chapter 4) in artificial intelligence, cognitive science, computer vision, and computational linguistics works within a probabilistic framework (Chater, Tenenbaum, & Yuille, 2006). The psychology of reasoning has widely adopted probabilistic ideas (e.g., Evans, Handley, & Over, 2003; McKenzie & Mikkelsen, 2007; Oaksford & Chater, 1994, 2007; see Buehner & Cheng, Chapter 12); and the same ideas have been extended to model argumentation (see Hahn & Oaksford, Chapter 15).

Rational Choice

From the point of view of the psychology of thought, logic can be viewed as providing consistency constraints on belief; and probability can be viewed as providing consistency constraints on degrees of belief. How far is it possible to specify consistency conditions over choices? This is the objective of the theory of rational choice (in this subsection, we draw some material from Allingham, 2002, who provides a beautifully concise introduction).

According to the theory, following Hume, no choice can be judged rational or irrational when considered in isolation. It may seem bizarre to choose a poke in the ribs, P , over a meal, M , in a top restaurant; but such a preference is not irrational.

But there surely is something decidedly odd about choosing P , when offered $\{P, M, D\}$; but choosing M , from the options $\{P, M\}$ where now a third “decoy” option has been removed. The first choice seems to indicate a preference for P over M ; but the second seems to indicate the contradictory preference of M over P . The *contraction* condition is that this pattern is disallowed.

Let us consider another possible rationality condition. Suppose that we choose P , when offered either $\{P, M\}$ or $\{P, D\}$, but we do not choose P when offered $\{P, M, D\}$ —perhaps we choose D . This again seems odd. The *expansion* condition is that this pattern is ruled out.

If we obey both the contraction and expansion conditions, there is a preference relation over our items, which we can interpret as meaning “at least as good as,” so that any choice from a set of options is at least as good as any of the other items (there may be more than one such choice). If this preference relation is *transitive* (i.e., X is at least as good as Y ; Y is at least as good as Z requires that X is at least as good as Z), then this preference relation is an ordering.

If these three consistency conditions are respected, it turns out that we can order each option (possibly with ties), with most favored options at one end, and least favored at the other. And, if we like, we can place the ordering on to the number line (in any way we like, as long as higher numbers represent items further to the favored end of the ordering)—it is conventional to call these numbers *utilities*. It is important to think of this notion as a purely formal notion, not necessarily connected with value or usefulness in any intuitive sense. So, we can now give a simple criterion for rational choice: When given a set of options, choose the item with the highest utility.

Here is a simple argument in favor of all three of these conditions: If a person violates them, it appears that the person’s money can systematically be taken from him or her. The person becomes a “money pump.” Suppose that agent, A , violates the contraction condition: that A chooses P , when offered $\{P, M, D\}$, but chooses M , from the options $\{P, M\}$. Money can be removed from this hapless agent by an evil counterparty, E , as follows. Suppose A , initially, has option P . E suggests that perhaps A might prefer the alternative M . Considering the set $\{P, M\}$, A prefers M , by hypothesis. This seems to imply that E can persuade A to “swap” to his or her preferred choice, by paying a sufficiently small sum of money, Δ_1 .

Now, having made the payment, A has M . Now the evil counterparty asks whether A might instead prefer either D or P . A now faces the choice set $\{P, M, D\}$. Now, by hypothesis, and in violation of the contraction condition, A prefers P . So E can persuade A to switch M for P , on payment of an arbitrarily tiny sum, Δ_2 . So the hapless A now has exactly the option he or she started with; but is $\Delta_1 + \Delta_2$ poorer. E repeats, removing A ’s money until there is none left. According to this argument, we violate the contraction condition at our peril.

So far we have focused on options with certain outcomes. But many choices are between options whose outcome is not certain. For example, we might wonder whether to buy a stock or make a gamble; or how much one would enjoy a particular choice on the menu. We now have some extra consistency conditions that seem natural. If I prefer P to M , then surely I should prefer a gamble with a probability q of P , and $1-q$ of X (for any option X), to a gamble with a probability q of M , and $1-q$ of X . This is the *substitution* condition.

And if I prefer P to R , and R to M , then it seems reasonable that there must be some s such that I am indifferent between R and a gamble mixing P and M , with probability s . After all, if $s = 1$, that is, P is certain, then the mixed gamble will be preferred to R ; when $s = 0$, and M is certain, the mixed gamble will be dispreferred. The *continuity* condition is simply that there must be some intermediate value of s where the mixed gamble is neither preferred nor dispreferred.

When and only when these two apparently mild conditions are respected, then the pattern of preferences over a gamble can be represented by a more precise “utility scale”—one which associates a real number with each outcome; the preferred option from some set is always that which has the highest expected utility (i.e., the average of the utilities of the possible outcomes, weighted by their probability). This is the normative principle of maximizing expected utility (EU).

EU has provided a fundamental starting point for many areas of economics; and EU has also been the standard against which the quality of human decision making is often judged. Crucially, from the point of view of this volume, EU has also been viewed as a starting point for *descriptive* theories of decision making, from which various departures can be made (see contributions in this volume by LeBoeuf and Shafir, Chapter 16; and Camerer & Smith, Chapter 18).

Exploring apparent departures from the predictions of rational choice theory and probability in human behavior has been a major field of research over the last 50 years—to a fair approximation, the entire field of judgment and decision making (e.g., Goldstein & Hogarth, 1997) is devoted to this topic. By contrast, rational choice theory provides a foundation for modern microeconomics (but see Camerer & Smith, Chapter 18); and the same “economic” style of explanation has been enormously productive in behavioral ecology (e.g., Stephens & Krebs, 1986).

Note, finally, that rational choice theory extends in a variety of directions that we have not touched on here. Perhaps the most important is *game theory*, the theory of rational strategic interaction between players, where the outcome of each agent’s actions depends on the action of the other. We will not treat this vast and enormously important topic here, as it arises infrequently throughout the rest of this volume. We have focused instead on thinking and reasoning within a single individual.

Conclusions

This volume is primarily concerned with the descriptive project of understanding how people think and reason. The present chapter, by contrast, has outlined normative theories that aim to say something about how people *should* think. More accurately, normative theories provide consistency conditions on thought. Logic can be viewed as helping to clarify which sets of beliefs are (in)consistent; probability theory can clarify which degrees of belief are consistent; and rational choice theory imposes consistency conditions on choices, values, and beliefs. Perhaps one reason why we think and reason at all is to attempt to reestablish such consistency, when it is disturbed—but the dynamic process of reestablishing equilibrium is relatively little understood.

But if we assume that the mind is at such an equilibrium, then the consistency conditions can be used to determine how people should believe or choose, given some given set of beliefs or choices. So, if a rational agent prefers *A* to *B*, and *B* to *C*, then the transitivity of preference requires that the agent prefers *A* to *C*. If a person believes *A* or *B*, and not *B*, then logical consistency (according to the standard translation into the propositional calculus) requires that the person believe *A*. This brings us back to the first of our three motivations for considering the relevance of normative theories to descriptive theories

of thought: that it describes the “right answers” in reasoning problems, just as arithmetic provides the right answers against which mental calculation can be judged. Note, though, that mere consistency conditions do not appear to provide the basis for an exhaustive analysis of the functions of thought. A variety of topics in the present volume, including the study of similarity (Goldstone & Son, Chapter 10), analogy (Holyoak, Chapter 13), creative thinking (Smith & Ward, Chapter 23), and insight (van Steenburgh et al., Chapter 24), seem not readily to be understood as merely avoiding inconsistency—and have largely been resistant to the encroachment of normative theory. Despite appearances, however, it remains possible that some of these aspects of thought may be governed by normative principles (e.g., Chater & Vitányi, 2003; Holyoak, Lee, & Lu, 2010; Tenenbaum & Griffiths, 2001; see also Holyoak, Chapter 13).

Second, note that consistency conditions provide the starting point for descriptive theories. One way in which human thinking and reasoning can adhere to the standards of a normative theory is by actually carrying out the calculations defined by the normative theory, at least to some approximation. Thus, mental logics and mental models are inspired by different approaches to logical proof; Bayesian cognitive science has been inspired by developments in probability theory; and most descriptive theories of decision making are departures, to varying degrees, from rational choice models.

Finally, picking up on our final role for normative accounts of rationality, we stress that without some normative constraints on thinking and reasoning, it becomes impossible to interpret thought, and resulting utterances or behavior, at all. We do not have direct access to people’s beliefs or degrees of belief; we have to infer them from what they say and how they behave (protestations that the food is safe to eat may be undermined by a stubborn refusal to eat any). And utilities (in the technical sense, recall) are defined by choices; but we can only observe a fraction of possible choices, and, as theorists, we have to infer the rest. Consistency conditions can, on this account, be critical in making such inferences possible, by binding together beliefs or choices that we have observed with beliefs and choices that we have not observed.

Future Directions

In this brief tour of normative theories of thinking and reasoning, we have, inevitably, focused on what

normative theories handle successfully. In closing, we highlight three areas that appear to provide challenges for current normative approaches.

The first area concerns understanding how thinking, reasoning, and decision making are influenced by world knowledge. One of the most important observations in early artificial intelligence and cognitive science research was the extraordinary richness of the knowledge required to understand even the simplest story or scenario (Clark, 1975; Minsky, 1977). Our thoughts effortlessly draw on rich knowledge of the physical and social worlds, not merely the logical forms of the sentences that we are hearing or reading; and such knowledge itself appears to have a “fractal” character (Chater & Oaksford, 2001). That is, explaining any given fact about the physical and social worlds appears to require drawing on yet further such knowledge, and so on indefinitely. While mathematical concepts, such as sets, groups, and the real line, can neatly be captured by a few axioms (although there is sometimes controversy about which axioms are most appropriate), real-world categories, such as “chair,” “country,” “person,” or “belief,” stubbornly resist such formalization (see Rips et al., Chapter 11). Rather, they appear to be part of an interdependent “web of belief” (Quine & Ullian, 1978), which is difficult, or perhaps even impossible, to characterize piecemeal. One consequence of this observation is that the specification of the world knowledge that underlies particular aspects of everyday reasoning will be difficult, if not impossible (c.f. Fodor’s, 1983, discussion of “central” cognitive processes; and AI skeptics such as Dreyfus, 1972). One particularly notorious problem that arises in this context is the “frame problem” (McCarthy & Hayes, 1969; Pylyshyn, 1987). Suppose that an agent decides to perform some action: The frame problem is to determine which aspects of the agent’s knowledge can be left unchanged, and which must be updated (by analogy with the question of which aspects of the background frame an animator can leave unchanged, or must modify from frame to frame, as a cartoon character moves). Despite appearing, at first sight, fairly innocuous, the frame problem was the rock upon which many proposals in early artificial intelligence foundered, in large part because it requires relating a local piece of new information (concerning the action) with the endless and ill-understood network of background world knowledge. A central task for future research is to establish how far it is possible to extend the use of current

methods based on normative principles to ever richer aspects of knowledge (e.g., Griffiths, Kemp, & Tenenbaum, 2008; Pearl, 2000) or to develop new tools to allow this.

A second challenging area for current normative models concerns dealing with inconsistency. We have stressed that normative models are typically inherently static: They determine what beliefs, degrees of belief, or beliefs, utilities, and actions are consistent with each other. But, on the face of it at least, human thought is riddled with inconsistency. We may believe that the probability of an air crash is infinitesimally low, but simultaneously be terrified of flying; we may be desperate to save for the future, yet spend compulsively; and, more prosaically, we may accept the axioms of arithmetic but believe that Fermat’s last theorem is false. Indeed, as this last case makes clear, the problem of determining whether your beliefs are consistent is enormously difficult. But if we accept that human thought is inconsistent, then we face two challenges. The first is finding general principles that determine how consistency should best be restored; and the second is avoiding the inferential chaos that may appear to result from the mere existence of inconsistency. This latter problem is particularly immediate in classical logic, in which a set of propositions S_1, S_2, \dots, S_n has, as a consequence proposition T unless it is possible that S_1, S_2, \dots, S_n are true, but T is false. If S_1, S_2, \dots, S_n are inconsistent, then these premises can never simultaneously be true, and hence this type of counterexample can be never generated. But this means that, from an inconsistency, anything at all follows. Inconsistencies propagate in similarly pathological ways in probability and rational choice theory. But if inconsistency is ubiquitous in human thought, then how can normative theories gain any explanatory purchase? One approach within logic is the development of so-called paraconsistent logics (Priest & Tanaka, 2009), which do not allow inferential “explosion” when a contradiction is reached. Nonetheless, the problem of dealing with the inconsistency of beliefs and choices remains extremely challenging for the application of normative theories to cognition.

Finally, following the dictates of a normative theory of reasoning precisely would require carrying out calculations of enormous complexity—for reasonably complex problems, such calculations appear to far exceed the capacity of human thought. For example, figuring out whether a particular set of beliefs is consistent, even in elementary logics such

as the propositional or first-order predicate calculus, is not, in general, computationally feasible (in the case of propositional calculus, the problem is NP-complete; see Cooke, 1971; for first-order logic it is undecidable; seeBoolos & Jeffrey, 1980). Such intractability is at least as troublesome, in general, for calculations concerning probability or rational choice (e.g., Binmore, 2008; van Rooij, 2008).

The problem of computational tractability is especially problematic for our second potential role for normative theories: providing the starting point for descriptive theories of thinking and reasoning. On the face of it, the mind cannot perfectly implement the precepts of logic, probability, or rational choice theory—because the mind is presumably limited to computable processes. One reaction to this problem is to entirely reject normative theory as a starting point for descriptive accounts. Gigerenzer and Goldstein (1996), for example, argue that rational choice theory requires that the mind is a Laplacian demon, with infinite computational resources; and argue instead that human judgment and decision making involves “fast and frugal” heuristics, unrelated to rational norms. Alternatively, perhaps human thought is a cheap *approximation* to normative calculations (Vul, Goodman, Griffiths, & Tenenbaum, 2009). More broadly, the relationship between normative and descriptive theories of thinking and reasoning is likely to remain an important area of controversy and future research.

References

- Allingham, M. (2002). *Choice theory: A very short introduction*. Oxford, England: Oxford University Press.
- Barwise, J., & Etchemendy, J. (2000). *Language, proof and logic*. Chicago, IL: University of Chicago Press.
- Binmore, K. (2008). *Rational decisions*. Princeton, NJ: Princeton University Press.
- Boolos, G. S., & Jeffrey, R. C. (1980). *Computability and logic* (2nd ed.). Cambridge, England: Cambridge University Press.
- Braine, M. D. S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review*, 85, 1–21.
- Chater, N., & Oaksford, M. (2001). Human rationality and the psychology of reasoning: Where do we go from here? *British Journal of Psychology*, 92, 193–216.
- Chater, N., Tenenbaum, J., & Yuille, A. (Eds.). (2006). Probabilistic models of cognition. Special Issue. *Trends in Cognitive Sciences*, 10 (whole issue).
- Chater, N., & Vitányi, P. (2003). The generalized universal law of generalization. *Journal of Mathematical Psychology*, 47, 346–369.
- Clark, H. H. (1975). Bridging. In R. C. Schank & B. L. Webber (Eds.), *Theoretical issues in natural language processing* (pp. 169–174). New York: Association for Computing Machinery.
- Cook, S. A. (1971). The complexity of theorem proving procedures. In *Proceedings of the Third Annual Association for Computing Machinery Symposium on Theory of Computing* (pp. 151–158). New York: Association for Computing Machinery.
- Cox, R. T. (1946). Probability, frequency, and reasonable expectation. *American Journal of Physics*, 14, 1–13.
- Davidson, D. (1984). *Inquiries into truth and interpretation*. Oxford, England: Oxford University Press.
- Dowty, D., Wall, R. E., & Peters, S. (1981). *Introduction to Montague semantics*. Dordrecht, Holland: Reidel.
- Dreyfus, H. (1972). *What computers can't do*, New York: Harper and Row.
- Edgington, D. (1995). On conditionals. *Mind*, 104, 235–329.
- Evans, J. St. B. T. (1984). Heuristics and analytic processes in reasoning. *British Journal of Psychology*, 75(4), 541–568
- Evans, J. St. B. T., Handley, S. J., & Over, D. E. (2003). Conditionals and conditional probability. *Journal of Experimental Psychology—Learning, Memory, and Cognition*, 29, 321–335.
- Evans, J. St. B. T., & Over, D. E. (2004). *If*. Oxford, England: Oxford University Press.
- Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, Scotland: Oliver & Boyd.
- Fodor, J. A. (1983). *Modularity of mind*. Cambridge, MA: MIT Press.
- Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28, 3–71.
- Gigerenzer, G., & Goldstein, D. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103, 650–669.
- Gigerenzer, G., & Todd, P. (Eds.). (1999). *Simple heuristics that make us smart*. Oxford, England: Oxford University Press.
- Goldstein, W. M., & Hogarth, R. M. (Eds.). (1997). *Judgment and decision making: Currents, connections, and controversies*. Cambridge, England: Cambridge University Press.
- Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge Handbook of Computational Psychology*. New York: Cambridge University Press.
- Harman, G. (1986). *Change in view*. Cambridge, MA: MIT Press.
- Holyoak, K. J., Lee, H. S., & Lu, H. (2010). Analogical and category-based inference: A theoretical integration with Bayesian causal models. *Journal of Experimental Psychological: General*, 139, 702–727.
- Inhelder, B., & Piaget, J. (1955). *De la logique de l'enfant à la logique de l'adolescent* [The growth of logical thinking from childhood to adolescence]. Paris: Presses Universitaires de France.
- Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, England: Cambridge University Press.
- Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Kolmogorov, A. N. (1956). *Foundations of the theory of probability* (2nd ed.). New York: Chelsea Publishing Company.
- Macnamara, J. (1986). *A border dispute: The place of logic in psychology*. Cambridge, MA: MIT Press.
- McCarthy, J., & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer & D. Michie (Eds.), *Machine intelligence 4* (pp. 463–502). Edinburgh: Edinburgh University Press.
- McKenzie, C. R. M., & Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology*, 54, 33–61.

- Minsky, M. (1977). Frame System Theory. In P. N. Johnson-Laird & P. C. Wason (Eds.), *Thinking: Readings in cognitive science* (pp. 355–376). Cambridge, England: Cambridge University Press.
- Nagel, E., & Newman, J. R. (1958). *Godel's proof*. New York: New York University Press.
- Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101, 608–631.
- Oaksford, M., & Chater, N. (2007). *Bayesian rationality*. Oxford, England: Oxford University Press.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Mateo, CA: Morgan Kaufmann.
- Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, England: Cambridge University Press.
- Priest, G., & Tanaka, K. (2009). Paraconsistent logic. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2009 ed.). Retrieved August 10, 2011, from <http://plato.stanford.edu/archives/sum2009/entries/logic-paraconsistent/>
- Pylyshyn, Z. (Ed.). (1987). *The robot's dilemma: The frame problem in artificial intelligence*. Norwood, NJ: Ablex.
- Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: Harvard University Press.
- Quine, W. V. O., & Ullian, J. S. (1978). *The web of belief*. New York: McGraw Hill.
- Ramsey, F. P. (1926). "Truth and probability." In R. B. Braithwaite (Eds.), *The foundations of mathematics and other logical essays* (pp. 156–198). London: Kegan Paul.
- Rips, L. J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.
- Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Stephens, D. W., & Krebs, J. R. (1986). *Foraging theory*. Princeton, NJ: Princeton University Press.
- Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24, 629–641.
- Tversky, A., & Kahneman, D. (1983). Extension versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90, 293–315.
- van Rooij, I.** (2008). The tractable cognition thesis. *Cognitive Science*, 32, 939–984.
- von Mises, R. (1939). *Probability, statistics, and truth*. New York: Macmillan.
- Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2009). One and done? Optimal decisions from very few samples. In *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society* (pp. 148–153). San Mateo, CA: Erlbaum.

Bayesian Inference

Thomas L. Griffiths, Joshua B. Tenenbaum, and Charles Kemp

Abstract

Inductive inferences that take us from observed data to underdetermined hypotheses are required to solve many cognitive problems, including learning categories, causal relationships, and languages. Bayesian inference provides a unifying framework for understanding how people make these inductive inferences, indicating how prior expectations should be combined with data. We introduce the Bayesian approach and discuss how it relates to other approaches such as the “heuristics and biases” research program. We then highlight some of the contributions that have been made by analyzing human cognition from the perspective of Bayesian inference, including connecting symbolic representations with statistical learning, identifying the inductive biases that guide human judgments, and forming connections to other disciplines.

Key Words: Bayesian inference, rational models, inductive inference, learning

Introduction

People solve a remarkable variety of problems in the course of everyday life, such as learning the categories that organize our world, identifying causal relationships, and acquiring language. Cognitive science has tended to take a “divide and conquer” approach to understanding the mind, analyzing each of these aspects of human learning separately. However, viewed abstractly, these problems all have something in common: They are inductive problems, requiring us to evaluate underdetermined hypotheses using limited data. The shared structure of these problems suggests that it might be possible to develop a single theory that would provide a unifying account of all of these different aspects of cognition.

Recently, a number of researchers have begun to explore the possibility that a unifying account of inductive inference might be provided by probability theory, as used in Bayesian statistics (Anderson, 1990; Chater & Oaksford, 1999; Oaksford & Chater, 1998; Tenenbaum, Griffiths, & Kemp, 2006). This is an idea that has been growing in

popularity in cognitive science, and it is bolstered by the widespread adoption of Bayesian methods in computer science (e.g., Bishop, 2006; Russell & Norvig, 2010) and statistics (e.g., Robert, 1994). Developing automated systems that can match human abilities in solving inductive problems such as learning categories, causal relationships, and languages has been a goal of artificial intelligence research for the last 50 years. Recently, significant steps toward this goal have been made by recognizing that ideas from Bayesian statistics can be useful in handling the uncertainty that is inherent in inductive inferences. Research in artificial intelligence and machine learning has led to powerful statistical models and efficient computer algorithms for working with those models.

Applying these computational and statistical ideas to human inductive inference has the potential to shed light on some of the deepest questions in cognitive science, such as the nature and origins of the constraints on human learning, and whether cognition is best thought about in terms of symbolic

representations or statistics. In this chapter, we review the basic ideas behind Bayesian inference, consider how this approach relates to other theoretical perspectives, and highlight some of the contributions that this approach has made to the investigation of human thinking and reasoning.

Basics of Bayesian Inference

If you were assessing the prospects of a 60-year-old man, how much longer would you expect him to live? If you heard that a movie had made \$40 million so far at the box office, how much would you expect it to make in total? Answering these questions requires an inductive inference that combines prior knowledge with the observed data. Bayesian inference provides a way to solve problems of this kind, and it relies upon one simple assumption: that our degrees of belief about a set of hypotheses can be expressed in terms of a probability distribution over those hypotheses. A variety of arguments have been given for this assumption, from proofs that it follows from a particular axiomatic description of common sense to proofs that a gambler who does not behave in this way will lose money (see Jaynes, 2003). However, in the context of cognitive science, we can adopt this as a methodological assumption—one that should be evaluated in terms of the amount of insight it ultimately gives us into human cognition.

If we are willing to express our degrees of belief in terms of probabilities, the problem of updating these beliefs in light of evidence is solved for us by the rules of probability theory. In particular, we should update our beliefs following a principle known as Bayes' rule. Assume that a learner has a prior probability distribution, $P(h)$, specifying the probability assigned to the truth of each hypothesis h in a set of hypotheses H (known as the hypothesis space) before seeing d . Bayes' rule states that the probability that should be given to each hypothesis after seeing d —known as the posterior probability, $P(h | d)$ —is

$$P(h | d) = \frac{P(d | h)P(h)}{\sum_{h \in H} P(d | h)P(h)} \quad (1)$$

where $P(d | h)$ —the likelihood—indicates how probable d is under hypothesis h .

We can apply Bayes' rule to any inductive problem. For example, when predicting human life spans, the hypotheses are total life spans and the

datum is the age of the man in question. The prior is the distribution over total life spans, as specified, for instance, by census data. The likelihood reflects the probability of encountering a man of a certain age given his total life span (e.g., an age greater than the total life span is impossible, so the probability of such an event would be zero). Combining prior and likelihood yields the posterior probability a learner should assign to each life span given the observed age, which can be used to make predictions. Griffiths and Tenenbaum (2006) used this analysis to study people's inductive predictions for a number of everyday events, including life spans and movie grosses, as well as the lengths of poems, political terms, or cake baking times. They showed a striking consistency between optimal Bayesian predictions and median human judgments, suggesting both that people have accurate knowledge of the distributions for these everyday events and that people can use this knowledge as Bayes' rule prescribes to predict new observed events (see Fig. 3.1).

Relationship to Other Theoretical Approaches

Bayesian models of cognition differ in some of their goals from other theoretical approaches to human cognition. In this section, we discuss the different levels at which thinking and reasoning can be analyzed, and how different theoretical approaches fit into this framework. We then consider how Bayesian models, which assume people express degrees of belief using probabilities and update these degrees of belief via mechanisms for probabilistic inference, can be reconciled with research in the “heuristics and biases” tradition that has famously argued against both of these assumptions.

Levels of Analysis

Marr (1982) argued that analyses of information processing systems can be conducted at three levels: the computational level, focusing on formalizing the problem being solved and identifying its ideal solution; the algorithmic level, focusing on the representations and processes that are used in executing that solution; and the implementation level, focusing on how these representations and processes are implemented in hardware. Several alternative frameworks for how to understand cognition at different levels of analysis have since been proposed (for a review, see Anderson, 1990), but all agree on the distinction between the functional characterization of a cognitive problem and its ideal solution, and

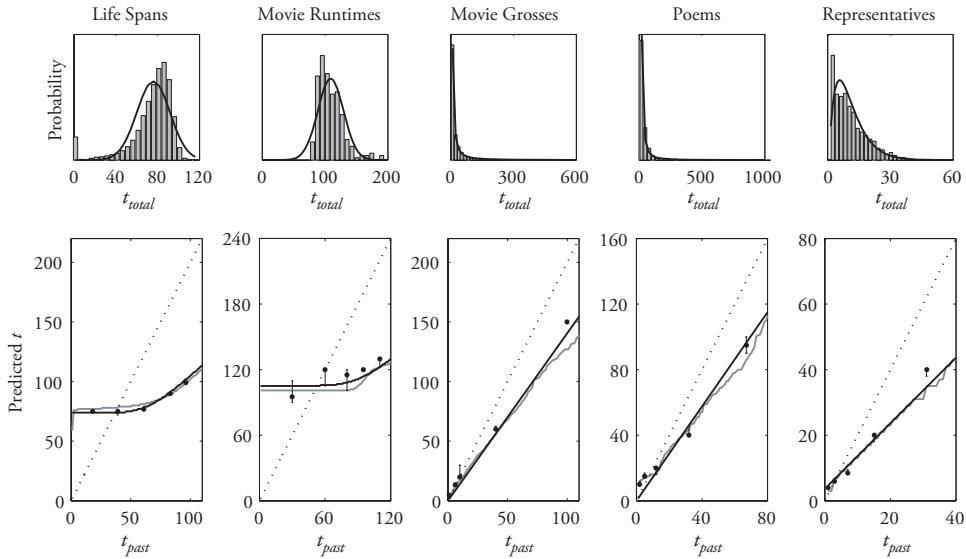


Fig. 3.1 People optimally integrate the evidence from a single observation with the appropriate prior distributions. The upper panels show the empirical distribution of the total duration or extent, t_{total} for five different everyday phenomena. The values of t_{total} are the hypotheses h to be evaluated, and these distributions are the appropriate priors. The first two distributions are approximately power-law, the next two are approximately Gaussian, and the last is approximately Erlang. Best-fitting parametric distributions are plotted in red. In the lower panels, black dots show subjects' median predictions for t_{total} when given a single observed sample t of a duration or extent in each of five domains (the data d used when applying Bayes' rule). Judgments are consistent with Bayesian predictions using the empirical prior distribution shown in the upper panel (black lines) and the best-fitting parametric prior (red lines). Predictions based on a single uninformative prior (dotted lines) are not consistent with these judgments. (Adapted from Griffiths and Tenenbaum, 2006.)

the mechanistic process that takes place inside people's heads.

The distinction between function and mechanism is important for understanding the goals of Bayesian models of cognition and how those goals differ from other theoretical perspectives. Bayesian inference could be taken as an account of human thinking and reasoning at either of these levels: as a proposed characterization of the ideal solution to the inductive problems that people face or as an account of the processes that people go through when they are solving those problems. However, most Bayesian models of cognition begin with and focus on the former, corresponding to Marr's (1982) computational level or Anderson's (1990) rational analysis. Many of the objections to Bayesian models of cognition result from assuming that these models are intended to apply straightforwardly or literally as mechanistic accounts, describing the actual processes that people follow when solving inductive problems.

In contrast to the "function first" approach to characterizing human inductive inference taken in many Bayesian models, most other approaches to computational modeling in psychology have emphasized the mechanistic level (see Doumas &

Hummel, Chapter 5). Symbolic cognitive architectures (e.g., Anderson, 1993) and connectionist models (e.g., McClelland & Rumelhart, 1986) are two primary examples. The components of these models, such as production rules or nodes in a neural network, are intended to correspond to isolable cognitive processes that support behavior and could in principle be reduced to some mechanism in the brain. If one looks at the basic ingredients of Bayesian models in the same way, as mechanistic proposals, even simple models could seem implausible. They would appear to require that human minds and brains explicitly represent very large, even infinite, hypothesis spaces, along with complex probability distributions over those spaces, and some kind of explicit implementation of Bayesian conditionalization. But this is not how most Bayesian models of cognition are intended to be interpreted, when framed at the computational level. They aim instead to identify the ideal solutions to inductive problems, given assumptions about the knowledge of the learner expressed through the choice of hypothesis space, prior, and likelihood. These models can be used to show that people act in a way that is at least approximately consistent with Bayesian

inference given hypothesis spaces, priors, and likelihoods of some form, providing insight into the kind of knowledge that guides people's inferences and constraints on the kind of algorithms that people could use to carry out those inferences.

Bayesian modelers are not uninterested in the mechanistic level, and they increasingly are turning their attention to integrating functional and mechanistic accounts—but still working within and extending the Bayesian paradigm. “Rational process models” attempt to characterize psychological mechanisms in terms of principled algorithms for approximate probabilistic inference, algorithms developed in Bayesian statistics and machine learning for approximate solution of large-scale problems that would be intractable for exact Bayesian calculations. Monte Carlo algorithms for sampling-based approximate inference have been particularly influential in recent work (Brown & Steyvers, 2009; Levy, Reali, & Griffiths, 2009; Sanborn, Griffiths, & Navarro, 2006; Vul, Goodman, Griffiths, & Tenenbaum, 2009) and computational neuroscientists have proposed that spiking in neural networks can implement forms of approximate probabilistic inference in biologically plausible ways (Fiser, Berkés, Orbán, & Lengyel, 2010). The Bayesian approach to cognitive modeling is thus best understood as a toolkit for reverse engineering the workings of human thought at all of Marr’s levels, following a top-down strategy that begins with ideal analyses at the computational level and then studies how these computations are implemented mechanistically at the algorithmic and hardware levels.

On Normativity, Reverse Engineering, and Heuristics

While Bayesian models emphasize ideal solutions to inductive problems, the way that they are used in cognitive modeling is different from traditional normative models. Logic, probability, and decision theory are often presented as defining a standard against which human thinking and reasoning can be compared (see Chater & Oaksford, Chapter 2). In such comparisons, the key question is whether people are rational, by the standards of those models, and in what ways they deviate from rational standards. In contrast, Bayesian models typically turn these questions around. These models ask what can we learn about human cognition by assuming people are rational, and what kind of rationality human thinking embodies (McKenzie, 2003). Rationality is thus not an empirical question, but

a methodological assumption that may or may not prove to be useful in understanding cognition.

Taking rationality as a guiding assumption derives directly from a focus on reverse engineering the mind, starting by viewing different aspects of cognition from a functional perspective. From an engineering standpoint, the central question is, “How does human cognition work so well, far better than any artificial intelligence system ever built?” Like any engineering solution, reverse-engineering accounts of human cognition should be framed in terms of rational solutions to the core problems, where “rational” does not mean ideal or globally optimal but simply “doing something for a reason,” which we can explore through a range of approaches to bounded or approximate or local optimization. The reverse-engineering questions then become, more precisely, “What is being optimized—what is the appropriate framing of the inductive problem to be solved?” and “How is it being optimized—what kind of approximations are used?”

The distinction between these two ways of using ideal models, the traditional normative approach and the reverse-engineering approach, arises starkly in what is often taken to be the greatest challenge to the view of cognition as Bayesian inference: the extensive literature on heuristics and biases that grew out of the work of Kahneman and Tversky (see Gilovich, Griffin, & Kahneman, 2002; Kahneman, Slovic, & Tversky, 1982; also see Frederick & Kahneman, Chapter 17). Their classic research program on judgment and decision under uncertainty set out to explore the cognitive processes by which people approach these problems that are normatively cast as probabilistic inference tasks, with an eye toward the heuristic rules people use to approximate rational probabilistic norms and the deviations from normative solutions (“biases”) that result from following these heuristics. Contemporary research in judgment and decision making (JDM) has focused on these deviations, leading to a view of human cognition as error prone and irrational (e.g., Ariely, 2008; Gilovich, 1993).

This JDM view has been the source of some natural and understandable skepticism about Bayesian approaches to cognitive modeling. Yet the Bayesian approach and the heuristics and biases program are more compatible than they might seem at first. Partly this is due to their complementary focus. Kahneman and Tversky, and much of the JDM work that followed, studied explicit reasoning about basic probability or statistics questions, often with

clear relevance to economic or policy questions, as well as more intuitive judgments about economic or policy questions where people's sense for the value of money is critical. People are typically asked to reason consciously and deliberately with information in more or less the same form that statisticians, policy makers, or economic decision makers typically work with: statistics, monetary values, and so on. It turns out that people often deviate from rational norms in these cases, and this has very important implications for many areas of applied decision making in our society.

This does not mean, however, that human minds always deviate from rational norms to such an extent or in these same ways. Much Bayesian modeling can be seen as an attempt to fill in this other side of the picture, to reverse engineer those aspects of human common sense that do work surprisingly well—better than any engineered machine reasoning systems. These are very different behavioral settings than typically studied in the JDM literature. The probabilistic judgments are often implicit, rather than explicit. The tasks are drawn less from the modern socioeconomic policy world and more from the real world of natural everyday experience—tasks which our brains have solved mostly unconsciously, automatically, and rapidly, in some form dating back to birth or early childhood. The data presented to people take the form of natural observations rather than explicitly stated statistics or monetary values. It turns out that in many of these settings, the principles of rational probabilistic inference are a valuable tool for explaining how our minds work—even the minds of young children (Perfors, Tenenbaum, Griffiths, & Xu, 2011).

Having said this, there are also significant areas of overlap in the JDM and Bayesian modeling literatures—both behavioral tasks and aspects of cognitive processing which both paradigms have studied and where it might seem that they have reached largely opposite conclusions (for example, the results of Griffiths and Tenenbaum, 2006, presented earlier seem inconsistent with previous work emphasizing errors in human predictions). We believe that many of these differences are superficial or can be resolved upon closer inspection. At heart we see real common purpose between the two approaches, although this can be hard for researchers in either camp to see with their different vocabularies and explanatory styles. We will try to draw out some of these connections in the following discussion.

When considering their implications for Bayesian models of cognition, the findings of nonnormative

behavior in the heuristics and biases research program can be divided into two broad classes. In the first class are results that can be viewed as rational, but under a different construal of the problem than the one the experimenter originally had in mind. Typically, the question of whether people are rational is framed by the experimenter defining a particular problem, and then examining whether people's behavior corresponds to a stipulated normative solution to that problem. However, there is no guarantee that people interpret the problem in the way intended by the experimenter or that the experimenter's interpretation is in any objective sense the "correct" one for the purposes of cognition in natural environments. If we instead assume that people are doing something reasonable, and consider what problem they might be solving, it often appears that people might be producing behavior that is a rational solution to a different problem—often a more complex, more interesting, but more natural problem. For example, Oaksford and Chater (1994) argued that apparently irrational behavior on Wason's (1966) card selection task could be understood as a rational attempt to maximize the information obtained by performing an action. Sher and McKenzie (2006) have conducted a number of experiments suggesting that framing effects and other instructional manipulations that are intended to change behavior without altering the form of a decision problem could result from a rational inference based on linguistic pragmatics. Griffiths and Tenenbaum (2001) argued that apparently irrational judgments about random processes—such as believing that HHTHT is more likely than HHHHH to be generated when a fair coin is flipped—could result from making a diagnostic inference about which outcome provides greater evidence in favor of a random process, rather than a predictive judgment about their probability under a random process. Likewise, our sense of coincidence seems to reflect the mirror image of subjective randomness as a rational inference: a measure of the strength of evidence for some structured hidden cause as the explanation for our observations, as opposed to a uniform random process (Griffiths & Tenenbaum, 2007a).

From a reverse-engineering perspective, these cases could be seen as signs that the computational-level analysis of the original experiments was miscast or inappropriate somehow. By showing how people's apparently irrational behavior can be interpreted as rational under an alternative framing, we come to a better understanding of the computational problem

the mind is actually solving. Of course, there are very real dangers of constructing “just so stories” here, so it is vital that these sorts of rational reanalyses be backed up by extensive and quantitative experimental support designed to test distinctive predictions of the proposed analysis versus compelling alternative or heuristic accounts. From a JDM viewpoint, these cases could be understood as people heuristically substituting an alternative problem or task that is easier for them to think about than the one originally posed (Kahneman & Frederick, 2002). This could lead to biases, and it could have serious policy or economic implications that no one could argue are rational. But this does not mean that the substituted problem, the heuristic shortcut, cannot itself be given a sensible rational analysis.

A second class of behavioral results are those that seem to admit no satisfying interpretation as Bayesian inference under a plausible problem formulation. These cases genuinely point to deviations from the ideal solutions to the problems people face and provide an opportunity to identify the cognitive mechanisms by which people might approximate ideal Bayesian computations. This style of explanation was arguably the intent of Kahneman and Tversky’s original work on heuristics and biases, although they did not have the theoretical vocabulary of approximate probabilistic inference that we do as a source of hypotheses for how cognitive mechanisms might work. Contemporary research in JDM focuses more on the errors that result from applying heuristics than on the value of those heuristics as approximations to computationally intractable inference problems. In some ways, this seems to reflect an expectation that human cognition should be error prone. For example, a very different view of the utility of ideal solutions can be found in research on human perception, where it is assumed that people do a good job of solving the inductive problems posed by the environment (Carpenter & Williams, 1995; Freeman, 1994; Huber, Shiffrin, Lyle, & Ruys, 2001; Kording & Wolpert, 2004; Platt & Glimcher, 1999; Weiss, Simoncelli, & Adelson, 2002). Consideration of inductive problems that go beyond the range typically considered in research on judgment and decision making, including problems such as categorization, causal learning, and language acquisition, reveals that people are generally very good at solving these problems. In fact, in many cases they are the only systems we know that can solve these problems, outperforming state-of-the-art machine learning systems. Viewed in this light, we

should be trying to understand how people succeed so often in solving these problems, rather than focusing on the cases where they make errors.

We should also comment briefly on how Bayesian models relate to another contemporary approach to JDM issues. In contrast to the emphasis on errors in much of the literature on judgment and decision making, Gigerenzer and colleagues (Gigerenzer et al., 1999) have argued that simple heuristics can in fact be adaptive. As part of this argument, they have shown that such heuristics can outperform more complex “rational” models, such as approaches to combining cue weights based on maximum-likelihood logistic regression. We do not view these results as being inconsistent with the claims typically made by Bayesian models—such heuristics might provide a way to approximate more computationally challenging probabilistic inference. However, the claim that the heuristics outperform rational models relies on comparing the heuristics against models that make weaker assumptions about the structure of the environment, in environments that match the assumptions made by the heuristics. The resulting advantage for the heuristics is simply an instance of what statisticians have termed the “bias-variance tradeoff,” with the more flexible model generalizing poorly (Geman, Bienenstock, & Doursat, 1992). When the rational models make assumptions about the environment more analogous to those made by the heuristics, they achieve similar or better performance (e.g., Martignon & Laskey, 1999).

Contributions to Understanding Thinking and Reasoning

Many of the central problems studied in cognitive science are inductive problems, including learning categories, identifying causal relationships, and acquiring language. Each of these problems can be analyzed using Bayesian inference. When learning categories, our hypotheses are ways of categorizing objects and our data are labeled category members (Anderson, 1990; Ashby & Alfonso-Reese, 1995; Sanborn, Griffiths, & Navarro, 2010). When learning causal relationships, our hypotheses are causal structures relating variables and our data are observations of those variables (Anderson, 1990; Griffiths & Tenenbaum, 2005). When acquiring language, our hypotheses are grammars or other schemes for expressing the probability of different utterances and our data are the utterances themselves (Chater & Manning, 2006; Goldwater, Griffiths, & Johnson, 2009; Perfors, Regier, & Tenenbaum, 2006). Viewing

all of these problems as different species of Bayesian inference has proven fertile in developing new cognitive models with a high degree of mathematical rigor, precision, and quantitative support in behavioral experiments. Rather than emphasizing the quantitative aspects of this work, here we describe the contributions in more conceptual terms—how Bayesian models enable cognitive scientists to think in new and productive ways about several questions of long-term interest.

Integrating Symbols and Statistics

Symbolic representation and statistical learning were long viewed as alternative or even opposing approaches to modeling the mind (Pinker, 1997). Models that rely on structured symbolic representations (e.g., Chomsky, 1957) have typically relied on weak learning mechanisms, and models that highlight the role of statistical learning (e.g., Rumelhart & McClelland, 1986) have typically relied on unstructured “subsymbolic” representations. Probabilistic models provide a new perspective on the debate between these approaches by showing how symbolic structures can be combined with statistical learning. Earlier statistical models of learning were typically defined over large arrays of numbers. For example, learning has been formulated as estimating strengths in an associative memory, weights in a neural network, or parameters of a high-dimensional nonlinear function (McClelland & Rumelhart, 1986; Rogers & McClelland, 2004). In contrast, Bayesian cognitive models often define probabilities over structured symbolic representations, including graphs, grammars, predicate logic, relational schemas, and functional programs.

In learning words and concepts from examples, the knowledge that guides both children’s and adults’ generalizations can be described using probabilistic models defined over tree-structured representations (Xu & Tenenbaum, 2007a, 2007b). Reasoning about other biological concepts for natural kinds—for example, given that cows and rhinos have protein X in their muscles, how likely is it that horses or squirrels do?—is also captured by Bayesian models that assume nearby objects in the tree are likely to share properties (Kemp & Tenenbaum, 2009). However, trees are by no means a “universal” representation. Inferences about other kinds of categories or properties are best captured using probabilistic models with different forms: two-dimensional spaces or grids for reasoning about geographic properties of cities, one-dimensional orders for

reasoning about values or abilities, or directed networks for causally transmitted properties of species (e.g., diseases) (Kemp & Tenenbaum, 2009).

Knowledge about causes and effects more generally can be expressed in a directed graphical model (Pearl, 1988): a graph structure where nodes represent variables and directed edges between nodes represent probabilistic causal links. In a medical setting, for instance, nodes might represent whether a patient has a cold, a cough, a fever, or other conditions, and the presence or absence of edges shows that colds tend to cause coughing and fever but not chest pain; lung disease tends to cause coughing and chest pain but not fever; and so on.

Such a “causal map” represents a simple kind of intuitive theory (Gopnik et al., 2004), but learning causal networks from limited data depends on the constraints of more abstract knowledge. For example, learning causal dependencies between medical conditions is enabled by a higher level framework theory (Wellman & Gelman, 1992) specifying two classes of variables (or nodes)—diseases (D) and symptoms (S)—and the tendency for causal relations (or graph edges) to run from D to S, rather than within these classes, or from S to D. This abstract framework can be represented using probabilistic models defined over relational data structures such as graph schemas (Kemp, Tenenbaum, Niyogi, & Griffiths, 2010), templates for graphs based on types of nodes, or probabilistic graph grammars (Griffiths & Tenenbaum, 2007b), similar in spirit to the probabilistic grammars for strings that have become standard for representing linguistic knowledge (Chater & Manning, 2006).

Models that rely on representations such as tree structures, causal maps, graph schemas, or graph grammars help to explain how inductive inferences are guided by different kinds of knowledge but also help to explain how these knowledge structures are acquired. If we take symbolic structures as our hypotheses, Bayes’ rule provides a way to learn which of these symbolic structures best describes our data. Combining structured representations and probabilistic inference therefore helps to explain how knowledge guides learning and how this knowledge is itself acquired.

Shedding Light on the Nature and Origins of Human Inductive Biases

Arguments from philosophy (Goodman, 1955) and formal analyses of learning (Geman et al., 1992; Kearns & Vazirani, 1994; Vapnik, 1995) indicate that the key to solving inductive problems is

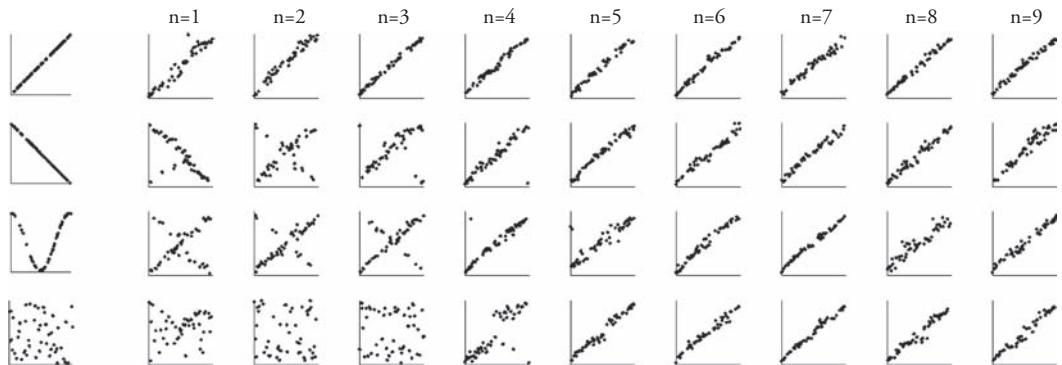


Fig. 3.2 Iterated learning reveals inductive biases. The leftmost panel in each row shows the set of samples from a function seen by the first learner in a sequence. The (x, y) pairs were presented as the lengths of two bars on a computer screen. During training, participants predicted the length of the y bar from the x bar, receiving feedback. The second panel shows the predictions produced by the first learner in a test phase, predicting y for a range of x values without feedback. These predictions were used as the training data for the second learner, who produced the predictions shown in the third column. The other panels show the data produced by each generation of learners, trained from the predictions produced by the previous learner. Regardless of the function used to generate the data shown to the first participant, iterated learning quickly converges to a linear function with positive slope, consistent with findings indicating that human learners are biased toward this kind of function. (Adapted from Kalish, Griffiths, & Lewandowsky, 2007.)

constraining the set of hypotheses under consideration. Bayesian methods provide a way to formalize the question of what kind of constraints guide human inductive inference and a language for expressing these constraints, through the prior distribution over hypotheses. They also allow the intermediate position that inferences can be strongly guided by knowledge acquired during a learner's lifetime, with priors being affected by the environment.

Bayesian models provide three ways to identify the constraints informing inductive inferences. The first, most obvious way is to conduct experiments and attempt to explain the results in terms of models using different prior distributions. This was the approach taken in the experiment by Griffiths and Tenenbaum (2006) presented earlier in the chapter, where the results suggested that people seemed to be using priors that corresponded well with the actual distributions of everyday quantities. The second way to identify constraints is to ask an abstract learnability question: What priors would be needed for people to learn the things that they do from the data that they observe? This kind of approach has typically been taken in studying language acquisition, where linguistic corpora provide a source of the utterances that children hear and can be used to obtain a good estimate of the data from which languages are learned (Goldwater et al., 2009; Perfors et al., 2006).

The third way to identify constraints on inductive inference is to specifically design experimental methods for estimating priors. One such method resulted from analyzing models of language evolution based

on “iterated learning”—the idea that every speaker of a language learns that language from another speaker, who had to learn it from somebody else in turn (Kirby, 2001). Formally, we can imagine a sequence of learners, each of whom receives data from the previous learner, forms a hypothesis from those data, and then uses that hypothesis to generate new data that are passed to the next learner. Griffiths and Kalish (2007) showed that if the learners use Bayes' rule, the probability that a learner selects a particular hypothesis converges to the prior probability assigned to that hypothesis as the length of the sequence increases. This result suggests a procedure for exploring human inductive biases: implement iterated learning in the laboratory with human learners and examine which hypotheses survive. Several experiments have shown that this method reveals constraints on learning consistent with the results of previous studies (Griffiths, Christian, & Kalish, 2006; Kalish, Griffiths, & Lewandowsky, 2007; Reali & Griffiths, 2009; see Fig. 3.2).

In addition to characterizing inductive constraints, psychologists attempt to understand the origin of these constraints. The acquisition of new inductive constraints is primarily the province of cognitive development (Carey, 2009; Gopnik & Meltzoff, 1997) but continues throughout life. In probabilistic terms, the challenge is to explain how hypothesis spaces and priors over these hypothesis spaces arise. Hierarchical Bayesian models (HBMs) (Gelman, Carlin, Stern, & Rubin, 1995) address the origins of hypothesis spaces and priors by positing

not just a single level of hypotheses to explain the data, but multiple levels: hypothesis spaces of hypothesis spaces, with priors on priors. Bayesian inference across all levels allows hypotheses and priors needed for a specific learning task to themselves be learned at larger or longer time scales, at the same time as they constrain lower level learning.

We illustrate the hierarchical Bayesian approach by building on the theme of the previous section and discussing how different kinds of symbolic representations can be learned. Symbolic representations provide a powerful source of constraints. For instance, in learning concepts over a domain of n objects, there are 2^n subsets and hence 2^n logically possible hypotheses for the extension of a novel concept. Assuming concepts correspond to the branches of a specific binary tree over the objects restricts this space to only $n - 1$ hypotheses. The constraints captured by a tree might be useful, but how might a learner find out which tree is best, and how might she discover in the first place that she should construct a tree rather than some other kind of representation? Discoveries like these have been pivotal in scientific progress: Mendeleev launched modern chemistry with his proposal of a periodic structure for the elements. Linnaeus famously proposed that relationships between biological species are best explained by a tree structure, rather than a simpler linear order (premodern Europe's "great chain of being") or some other form. Such structural insights have long been viewed by psychologists and philosophers of science as deeply mysterious in their mechanisms—more magical than computational. Conventional algorithms for unsupervised structure discovery in statistics and machine learning—hierarchical clustering, principal components analysis, multidimensional scaling, clique detection—assume a single fixed form of structure (Shepard, 1980). Unlike human children or scientists, they cannot learn multiple forms of structure or discover new forms in novel data.

Recently cognitive modelers have begun to explore how hierarchical Bayesian models can address these challenges. Kemp and Tenenbaum (2008, 2009) showed how HBMs defined over graph- and grammar-based representations can discover the form of structure governing similarity in a domain. Structures of different forms—trees, clusters, spaces, rings, orders, and so on—can all be represented as graphs, while the abstract principles underlying each form are expressed as simple grammatical rules for growing graphs of that form. Embedded in a hierarchical Bayesian framework, this approach can discover

the correct forms of structure (the grammars) for many real-world domains, along with the best structure (the graph) of the appropriate form, as shown in Figure 3.3.

Hierarchical Bayesian models can also be used to learn abstract causal knowledge, such as the framework theory of diseases and symptoms, and other simple forms of intuitive theories (Kemp et al., 2010). Mansinghka et al. (2006) showed how a graph schema representing two classes of variables, diseases (D) and symptoms (S), and a preference for causal links running from D to S nodes, can be learned from the same data that support learning causal links between specific diseases and symptoms—and learned just as fast or faster (see Fig. 3.4). The learned schema, in turn, dramatically accelerates learning of specific causal relations at the level below. Getting the big picture first—discovering that diseases cause symptoms before pinning down any specific disease-symptom link—and then using that framework to fill in the gaps of specific knowledge is a distinctively human mode of learning. It figures prominently in children's development and scientific progress but has not previously fit into the landscape of rational or statistical learning models.

Across several case studies of learning abstract knowledge—discovering structural forms, causal framework theories, and other inductive constraints acquired through transfer learning—it has been found that abstractions in HBMs can be learned remarkably fast, from relatively little data compared to what is needed for learning at lower levels. This is because each degree of freedom at a higher level of the HBM influences—and pools evidence from—many variables at levels below. This property of HBMs has been called the "the blessing of abstraction" (Goodman, Ullman, & Tenenbaum, 2011). It offers a top-down route to the origins of knowledge that contrasts sharply with the two classic theories of development: nativism (Chomsky, 1986; Spelke, Breinlinger, Macomber, & Jacobson, 1992), in which abstract concepts must be present from birth, and empiricism or associationism (Rogers & McClelland, 2004), in which abstractions are constructed but only slowly, in a bottom-up fashion, by layering many experiences on top of each other and filtering out their common elements. Only HBMs thus seem suited to explaining the two most striking features of abstract knowledge in humans: that it can be learned from experience, and that it can be present remarkably early in life, serving to constrain more specific learning tasks.

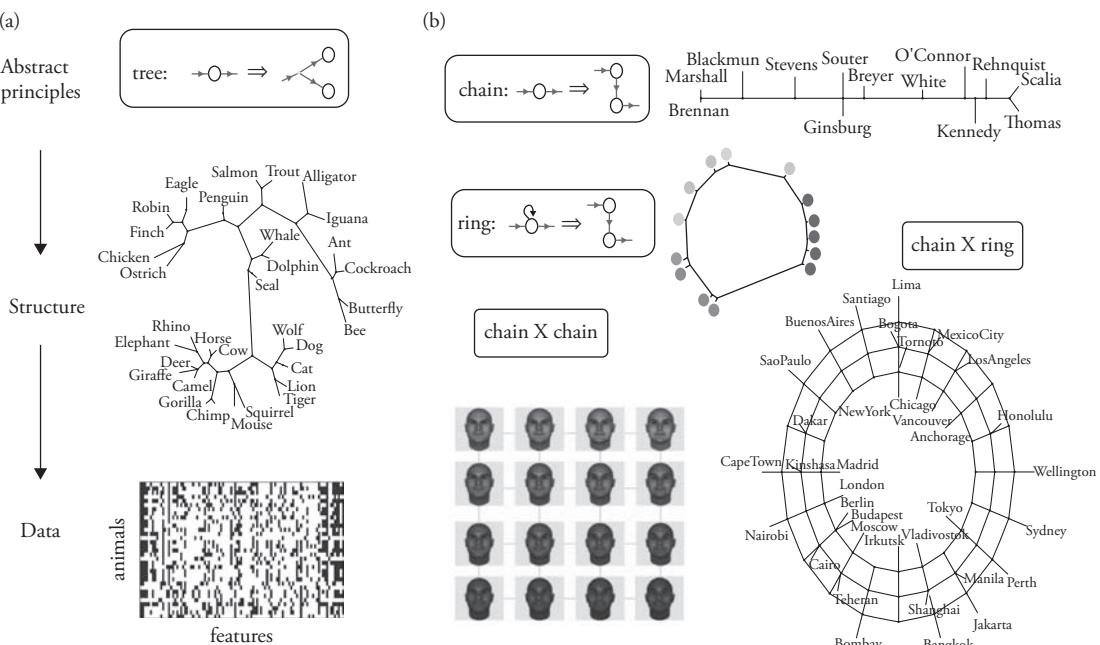


Fig. 3.3 Kemp and Tenenbaum (2008) showed how the form of structure in a domain can be discovered using a hierarchical Bayesian model defined over graph grammars. At the bottom level of the model is a data matrix D of objects and their properties, or similarities between pairs of objects. One level up is a graph describing how properties are distributed over objects. Intuitively, objects nearby in the graph are expected to share properties. At the highest level grammatical rules specify the form of structure in the domain—rules for growing graphs of a constrained form out of an initial seed node. A search algorithm attempts to find the combination of a form grammar F and graph G generated by that grammar which jointly receive highest posterior probability $P(F,G|D)$. (a) Given observations about the features of animals, the algorithm infers that a tree structure best explains the data. The best tree found captures intuitively sensible categories at multiple scales. (b) The same algorithm discovers that the voting patterns of U.S. Supreme Court judges are best explained by a linear “left-right” spectrum, and that subjective similarities among colors are best explained by a circular ring. Given images of faces varying in two dimensions, race and masculinity, the algorithm successfully recovers the underlying two-dimensional grid structure. Given proximities between cities on the globe, the algorithm discovers a cylindrical representation analogous to latitude and longitude. See color figure.

Developing Connections to Other Disciplines

Considering the ideal solutions to inductive problems has often resulted in surprising connections to ideas in other disciplines. In particular, statistics, artificial intelligence, and machine learning are disciplines that all involve inductive problems, and all are interested in rational solutions to those problems. Probabilistic approaches have recently become the standard approach to induction in all of these disciplines. If people are solving inductive problems well, we might expect that this lingua franca extends to psychology. To the extent that the inductive problems being addressed in psychology and the engineering disciplines take the same form, with the same inputs and outputs and the constraints of operating in the same world, we might even expect to find precise correspondences between the solutions adopted by the human mind and

those that have been identified in these other fields. For example, Ashby and Alfonso-Reese (1995) observed that a number of psychological models of categorization, such as exemplar and prototype models, can be viewed as corresponding to different methods of estimating a probability distribution, a problem known as “density estimation” in statistics. Likewise, Anderson’s (1990, 1991) rational model of categorization corresponds to a class of models used in nonparametric Bayesian statistics (Neal, 1998). This connection provides a link to a rich literature building on this class of models, which can be a source of new psychological hypotheses (Sanborn et al., 2010).

Recognizing connections to other disciplines is also useful because mathematical tools from those disciplines can sometimes be used to clarify psychological questions. Recent work on causal induction provides a prime example. Griffiths and Tenenbaum

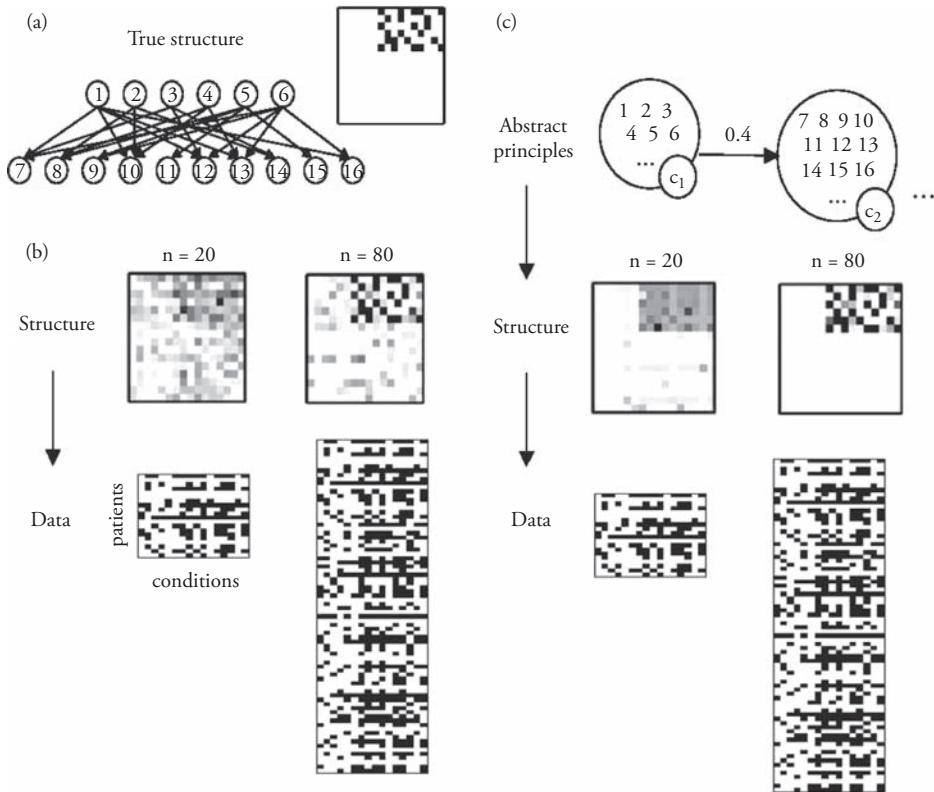


Fig.3.4 Hierarchical Bayesian models defined over graph schemas can acquire intuitive theories and use those theories to learn about specific causal relations from limited data (Kemp, Tenenbaum, Niyogi, & Griffiths, 2010). (a) For example, a simple medical reasoning domain might be described by relations among 16 variables: the first six are “diseases,” with causal links to the next ten “symptoms.” This causal network can be visualized as a binary matrix (upper right). (b) A two-level hierarchical Bayesian model (HBM) formalizes bottom-up causal learning, or learning with an uninformative prior on networks. At bottom we observe data D on patients and the conditions afflicting them. The second level encodes causal relations among these conditions; a grayscale matrix visualizes the posterior probability that any given pairwise causal link exists, conditioned on observing n patients. The true causal network can be recovered perfectly only from observing very many patients ($n = 1,000$; not shown). With $n = 80$, spurious links are inferred, and with $n = 20$ almost nothing can be learned. (c) Mansinghka et al. (2006) introduced a three-level HBM, where the highest level is a graph schema. The schema can represent the disease-symptom theory by assigning nodes to classes C1 and C2, and placing a prior on the level below favoring C1 → C2 links. The effective number of node classes is discovered automatically via the Bayesian Occam’s razor. Although this three-level model has many more degrees of freedom than the model in (b), causal learning is faster and more accurate. With $n = 80$ data points, the causal network is identified near-perfectly. Only $n = 20$ data points are sufficient to learn the high-level C1 → C2 schema and thereby to limit uncertainty at the network level to just the question of which diseases cause which symptoms.

(2005) used causal graphical models, a formal language for representing and reasoning about causal relationships that was developed in computer science and statistics (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993), to make the distinction between learning whether a causal relationship exists and estimating the strength of this relationship. This distinction was valuable because existing rational models could be shown to correspond to methods for estimating the strength of a relationship, while some of the judgments that people made in causal induction tasks seemed better characterized in terms of inferences about whether a relationship existed.

Subsequent work has built on this observation, producing a more complete Bayesian account of how people might estimate the strength of causal relationships (Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008). Developmental psychologists have shown how Bayesian inference over appropriate hypothesis spaces of graphical models can illuminate children’s causal learning (Gopnik et al., 2004; Griffiths & Tenenbaum, 2007b; Sobel, Tenenbaum, & Gopnik, 2004). Other areas of perception and cognition that were not always seen as species of causal inference, such as classical conditioning, multisensory integration, or perceptual organization, have been

influenced by this work and usefully informed by a new generation of causal graphical models (Courville, Daw, & Touretzky, 2006; Kording et al., 2007; Orbán, Fiser, Aslin, & Lengyel, 2008).

The work on rational process models described earlier, taking Bayesian approaches down to the algorithmic and hardware levels, represents a particularly fertile flow of ideas between the engineering disciplines and psychology. To date, the idea flow has been mostly from the engineering side into psychology, with various principled Monte Carlo methods for approximate inference suggesting new rational interpretations of classic cognitive processing mechanisms, as well as new processing accounts. But cognitive scientists thinking about the challenges of developing more flexible and robust approximate inference schemes have also made contributions back to the engineering side. For example, Vul et al. (2009) show when and why decision making based on just a single sample from the Bayesian posterior may represent an optimal tradeoff between the expected gain of a correct decision and the expected reward timescale. These analyses could be useful in understanding how very efficient sampling-based approximate inference might work in the brain, or how it could most efficiently be instantiated in robots. Goodman, Mansinghka, Roy, Bonawitz, and Tenenbaum (2008) have proposed and implemented a general-purpose approach to Monte Carlo inference on arbitrary computable probability distributions, or probabilistic programs, which was originally developed for modeling process-level effects in rule-based categorization (Goodman, Tenenbaum, Feldman, & Griffiths, 2008). Much of the work described earlier in building rich hierarchical or nonparametric models for learning categories and causal networks can also be seen as importing ideas from machine learning and computer science into psychology. But again, cognitive scientists have built on these first generations of probabilistic models and developed new artificial intelligence and machine learning techniques motivated by trying to capture distinctively human learning abilities, such as the infinite relational model for learning simple relational theories (Kemp et al., 2010), the hierarchical Bayesian approach to structural form discovery (Kemp & Tenenbaum, 2008), or nonparametric Bayesian models for feature discovery (Griffiths & Ghahramani, 2006).

Conclusions and Future Directions

The Bayesian toolkit offers several contributions to understanding human thinking and reasoning. It

provides a unifying mathematical language for framing cognition as the solution to inductive problems and for building principled quantitative models of thought with a minimum of free parameters and ad hoc assumptions. Deeper, it offers a framework for understanding why the mind works the way it does, in terms of rational inference adapted to the structure of real-world environments, and what the mind knows about the world—abstract schemas and intuitive theories revealed only indirectly through how they constrain generalizations. The new tools we obtain by adopting this perspective allow us to integrate symbolic representations with statistical learning, identify human inductive biases and understand their origins, and connect cognitive psychology with other scientific disciplines.

Most important, the Bayesian approach lets us move beyond classic “either-or” dichotomies that have long shaped and limited debates in cognitive science: “empiricism versus nativism,” “domain-general versus domain-specific,” “logic (rules, symbols) versus probability (statistics, numbers).” Instead we can ask harder questions of reverse engineering, with answers potentially rich enough to help us build more human-like artificial intelligence. The future directions for Bayesian models involve grappling with some of the central questions of cognitive science. How can domain-general mechanisms of learning and representation build domain-specific systems of knowledge? How can structured symbolic knowledge be acquired by statistical learning? The answers that are emerging from current research suggest new ways to think about the development of a cognitive system. Powerful abstractions can be learned surprisingly quickly, together with or prior to learning the more concrete knowledge they constrain. Structured symbolic representations need not be rigid, static, hardwired, or brittle. Embedded in a probabilistic framework, they can grow dynamically and robustly in response to the sparse, noisy data of experience.

Acknowledgments

The writing of this chapter was supported in part by grants from the James S. McDonnell Foundation Causal Learning Research Collaborative (TLG and JBT), the Air Force Office of Scientific Research (TLG, grant number FA-9550-10-1-0232), and the National Science Foundation (TLG, grant number IIS-0845410).

References

- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.
- Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98(3), 409–429.
- Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.

- Ariely, D. (2008). *Predictably irrational: The hidden forces that shape our decisions*. New York: Harper Collins.
- Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, 39, 216–233.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.
- Brown, S. D., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology*, 58, 49–67.
- Carey, S. (2009). *The origin of concepts*. New York: Oxford University Press.
- Carpenter, R. H. S., & Williams, M. L. L. (1995). Neural computation of log likelihood in the control of saccadic eye movements. *Nature*, 377, 59–62.
- Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, 10, 335–344.
- Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Science*, 3, 57–65.
- Chomsky, N. (1957). *Syntactic structures*. The Hague, Netherlands: Mouton.
- Chomsky, N. (1986). *Language and problems of knowledge: The Managua lectures*. Cambridge, MA: MIT Press.
- Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, 10, 294–300.
- Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Sciences*, 14, 119–130.
- Freeman, W. T. (1994). The generic viewpoint assumption in a framework for visual perception. *Nature*, 368, 542–545.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. New York: Chapman & Hall.
- Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias-variance dilemma. *Neural Computation*, 4, 1–58.
- Gigerenzer, G., Todd, P. M., & the ABC Research Group (1999). *Simple heuristics that make us smart*. Oxford, England: Oxford University Press.
- Gilovich, T. (1993). *How we know what isn't so: The fallibility of reason in everyday life*. New York: Free Press.
- Gilovich, T., Griffin, D. W., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge, England: Cambridge University Press.
- Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112, 21–54.
- Goodman, N. (1955). *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.
- Goodman, N. D., Mansinghka, V. K., Roy, D. M., Bonawitz, K., & Tenenbaum, J. B. (2008). Church: A language for generative models. In *Proceedings of the 24th Annual Conference on Uncertainty in Artificial Intelligence* (UAI 24) (pp. 220–229). Corvallis, OR: AUAI Press.
- Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, 32, 108–154.
- Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, 118, 110–119.
- Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111, 1–31.
- Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.
- Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2006). Revealing priors on category structures through iterated learning. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1394–1399). Mahwah, NJ: Erlbaum.
- Griffiths, T. L., & Ghahramani, Z. (2006). Infinite latent feature models and the Indian buffet process. In *Advances in neural information processing systems 18* (pp. 475–482). Cambridge, MA: MIT Press.
- Griffiths, T. L., & Kalish, M. L. (2007). A Bayesian view of language evolution by iterated learning. *Cognitive Science*, 31, 441–480.
- Griffiths, T. L., & Tenenbaum, J. B. (2001). Randomness and coincidences: Reconciling intuition and probability theory. In *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society* (pp. 370–375). Hillsdale, NJ: Erlbaum.
- Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, 51, 354–384.
- Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, 17, 767–773.
- Griffiths, T. L., & Tenenbaum, J. B. (2007a). From mere coincidences to meaningful discoveries. *Cognition*, 103, 180–226.
- Griffiths, T. L., & Tenenbaum, J. B. (2007b). Two proposals for causal grammars. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 323–345). Oxford, England: Oxford University Press.
- Huber, D. E., Shiffrin, R. M., Lyle, K. B., & Ruys, K. I. (2001). Perception and preference in short-term word priming. *Psychological Review*, 108, 149–182.
- Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge, England: Cambridge University Press.
- Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (p. 49–81). Cambridge, England: Cambridge University Press.
- Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, England: Cambridge University Press.
- Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin and Review*, 14, 288–294.
- Kearns, M., & Vazirani, U. (1994). *An introduction to computational learning theory*. Cambridge, MA: MIT Press.
- Kemp, C., & Tenenbaum, J. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences USA*, 105, 10687.
- Kemp, C., & Tenenbaum, J. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, 116(1), 20–58.
- Kemp, C., Tenenbaum, J., Niyogi, S., & Griffiths, T. (2010). A probabilistic model of theory formation. *Cognition*, 114, 165–196.
- Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Journal of Evolutionary Computation*, 5, 102–110.

- Körding, K., Beierholm, U., Ma, W. J., Tenenbaum, J. B., Quartz, S., & Shams, L. (2007). Causal inference in multi-sensory perception. *PLoS ONE*, 2, e943.
- Körding, K., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, 427, 244–247.
- Levy, R., Reali, F., & Griffiths, T. L. (2009). Modeling the effects of memory on human online sentence processing with particle filters. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems 21* (pp. 937–944).
- Lu, H., Yuille, A., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, 115, 955–984.
- Mansinghka, V. K., Kemp, C., Tenenbaum, J. B., & Griffiths, T. L. (2006). Structured priors for structure learning. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI)* (pp. 324–331). Corvallis, OR: AUAI Press.
- Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.
- Martignon, L., & Laskey, K. (1999). Bayesian benchmarks for fast and frugal heuristics. In G. Gigerenzer, P. M. Todd & The ABC Research Group (Eds.), *Simple heuristics that make us smart* (p. 169–188). Oxford, England: Oxford University Press.
- McClelland, J., & Rumelhart, D. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, MA: MIT Press.
- McKenzie, C. R. M. (2003). Rational models as theories—not standards—of behavior. *Trends in Cognitive Sciences*, 7, 403–406.
- Neal, R. M. (1998). *Markov chain sampling methods for Dirichlet process mixture models* (Tech. Rep. No. 9815). Department of Statistics, University of Toronto, Toronto, ON.
- Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101, 608–631.
- Oaksford, M., & Chater, N. (Eds.). (1998). *Rational models of cognition*. Oxford, England: Oxford University Press.
- Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences USA*, 105, 2745–2750.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco, CA: Morgan Kaufmann.
- Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, England: Cambridge University Press.
- Perfors, A., Regier, T., & Tenenbaum, J. B. (2006). Poverty of the stimulus? A rational approach. In *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society* (pp. 663–668). Hillsdale, NJ: Erlbaum.
- Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, 120, 302–321.
- Pinker, S. (1997). *How the mind works*. New York: Norton.
- Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, 400, 233–238.
- Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, 111, 317–328.
- Robert, C. P. (1994). *The Bayesian choice: A decision-theoretic motivation*. New York: Springer.
- Rogers, T., & McClelland, J. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.
- Rumelhart, D., & McClelland, J. (1986). On learning the past tenses of English verbs. In J. McClelland, D. Rumelhart, & The PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2, pp. 216–271). Cambridge, MA: MIT Press.
- Russell, S. J., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.
- Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 726–731). Hillsdale, NJ: Erlbaum.
- Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, 117, 1144–1167.
- Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, 210, 390–398.
- Sher, S., & Mckenzie, C. R. M. (2006). Information leakage from logically equivalent frames. *Cognition*, 101, 467–494.
- Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science*, 28, 303–333.
- Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, 99, 605–632.
- Spirites, P., Glymour, C., & Scheines, R. (1993). *Causation prediction and search*. New York: Springer-Verlag.
- Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Science*, 10, 309–318.
- Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York: Springer.
- Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2009). One and done? optimal decisions from very few samples. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 148–153). Austin, TX: Cognitive Science Society.
- Wason, P. C. (1966). Reasoning. In B. Foss (Ed.), *New horizons in psychology* (pp. 135–151). Harmondsworth, England: Penguin.
- Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, 5, 598–604.
- Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, 43, 337–375.
- Xu, F., & Tenenbaum, J. B. (2007a). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, 10, 288–297.
- Xu, F., & Tenenbaum, J. B. (2007b). Word learning as Bayesian inference. *Psychological Review*, 114, 245–272.

Knowledge Representation

Arthur B. Markman

Abstract

Theories in psychology make implicit or explicit assumptions about the way people store and use information. The choice of a format for knowledge representation is crucial, because it influences what processes are easy or hard for a system to accomplish. In this chapter, I define the concept of a representation. Then, I review three broad types of representations that have been incorporated into many theories. Finally, I examine proposals for the role of body states in representation as well as proposals that the concept of knowledge representation has outlived its usefulness.

Key Words: knowledge representation, features, structured representation, analogy, embodied cognition, dynamical systems, multidimensional scaling

Introduction

Theories of psychological functioning routinely make assumptions about the type of information that people use to carry out a process. They also make proposals for the form in which that information is stored and the procedures by which it is used. The type, form, and use of information by psychological mechanisms are incorporated into psychological proposals about *knowledge representation*. This chapter aims to provide a broad introduction to knowledge representation (see Markman, 1999, for a more detailed discussion of these issues).

In this chapter, I start by defining what I mean by a representation. Then, I discuss some of the kinds of information that people have proposed to be central to people's mental representations. Next, I describe three proposals for the way that people's knowledge is structured. Finally, I explore some broader controversies within the field. For instance, I describe an antirepresentationalist argument that suggests that we can safely dispense with the notion

of knowledge representation. I end with a call for a pluralist approach to representation.

Mental Representations

The modern notion of a mental representation emerged during the *cognitive revolution* of the 1950s, when the computational view of the mind ascended. The behaviorist approach to psychology that played a significant role in American psychology explicitly denied that the form and content of people's knowledge were legitimate objects of scientific study. Advances in the development of digital computers, however, provided a theoretical basis for thinking about how information could be stored and manipulated in a device.

On this computational view, *minds* are descriptions of the nature of the program that is implemented (in humans) by *brains*. Just as computers with very different hardware architectures could implement the same word-processing program (and thus make use of the same data structures and algorithms), different brains might compute the same sorts of functions and thus

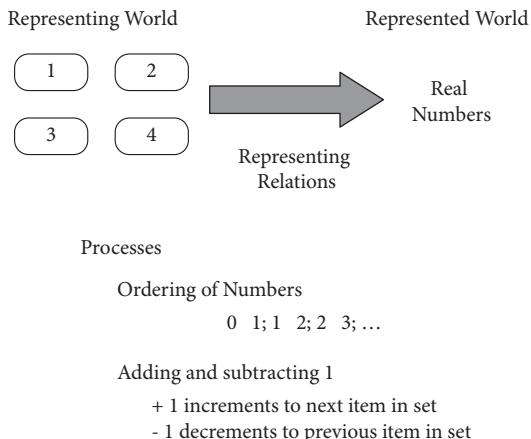


Fig. 4.1 Example of the four key aspects of a definition of a representation. A representing world corresponds to a represented world through some set of representing relations. Processes must be specified that make use of the information in the representation.

create representations (mental data structures) and processes (mental algorithms). Thus, proposals for knowledge representation are often stated at a level of description that abstracts across the details of what the brain is doing to implement that process yet nonetheless specifies in some detail how some functions are computed. Marr (1982) called this the *algorithmic level* of description. He called the abstract function being computed by an algorithm the *computational level* (see Griffiths, Tenenbaum, & Kemp, Chapter 3). What the brain is doing, which Marr (1982) called the *implementational level* of description, may provide some constraints on our understanding of these mental representations (see Morrison & Knowlton, Chapter 6; Green & Dunbar, Chapter 7), but not all of the details of an implementation are necessary to understand the way the mind works.

Defining Mental Representation

To define the concept of representation, I draw from work by Palmer (1978) and Pylyshyn (1980). In order for something to qualify as a representation, four conditions have to hold. First, there has to be some *representing world*. The representing world is the domain serving as the representation. In theories of psychological processing, this representing world may be a collection of symbols or a multidimensional space. In Figure 4.1, mental symbols for the numbers 0–9 are used to represent numerical quantities.

There is also a *represented world*. The represented world is the domain or information that is to be represented. There is almost always more information

in the world than can be captured by some representing world. Thus, the act of representing some world almost always leads to some loss of fidelity (Holland, Holyoak, Nisbett, & Thagard, 1986). In the example in Figure 4.1, representing the world of numbers with mental symbols will lose information, because the space of numbers is continuous, while this symbolic representation has discrete elements in it.

The third component of a representation is a set of *representing relations* that determines how the represented world stands in for the represented world. In Figure 4.1, these representing relations are shown as an arrow connecting the representing world to the represented world. There has to be some kind of relationship between the items that are used as the representation and the information being represented. These relations are what give the representation its meaning. I'll discuss this issue in more detail in the next section.

Finally, in order for something to function as a representation, there has to be some set of processes that use the information in the representation for some function. Without some processes that act on the representing world, the potential information in the representation is inert. In the case of the number representation in Figure 4.1, there need to be procedures that generate an ordering of the numbers and procedures that specify operations such as how to add or multiply or create other functions defined over numbers.

One reason why it is important to specify both the represented world and the processes is that it is tempting for readers to look at a representing world and have intuitions about the information that is captured in it. For example, you probably know a lot about numbers, and so you may bring that knowledge to bear when you see that the symbols in Figure 4.1 are associated with numbers. However, unless the system that is going to use this representation has procedures for manipulating these symbols, then the system that has this representing world does not have the same degree of knowledge that you do.

The definition that I just gave does not distinguish between mental representations and other things that serve as representations. For example, I may take a digital photograph as a representation of a visual scene. The photograph represents color by having a variety of pixels bunched close together that take on the wavelength of light that struck a detector when the picture was taken. This lawful

relationship is what allows this photo to serve as a representation of some scene. Furthermore, when you look at that photo, your visual and conceptual systems are able to interpret what is in the image, thereby serving as a process for extracting and using information from the representation.

In order for something to count as a mental representation, of course, the representing world needs to be inside the head. Humans use a variety of external representations, including photos, writing, and a variety of tools. The relationship between humans and the representations they put in their environment is interesting, and I discuss it a bit more at the end of this chapter.

Giving Meaning to Representations

In a computer program, data structures function to store the data that a program is going to manipulate in order to carry out some function. As I discussed earlier, a mental representation plays a role within cognitive systems that is analogous to the role of the data structure in a program. In order for this representation to serve effectively, though, there has to be some way to ensure that the representation captures the right information. We can think of the information that is captured by a particular representation as its *meaning*.

What are the sources of meaning in mental representations? Obviously, one source of meaning is the set of representing relations that relate the representing world to the represented world. These relations help to *ground* the representation. As a very simple example, consider a thermostat. One way to build a thermostat is to use a bimetal strip. The two metals that make up the strip expand and contract at different rates when exposed to heat. Thus, the curvature of the strip changes continuously with changes in temperature. In a typical thermostat, the bimetal strip is connected to a vial of conducting liquid (like mercury) that will close an electrical switch between two contacts when the right level of curvature is reached. Within this system, the lawful relationship between air temperature and the curvature of the strip provides a grounding for the representation of temperature.

Many mental representations that get (part of) their meaning through grounding are actually grounded in other representations. For example, research on vision focuses on how people detect the edges of objects in an image. Even these basic processes, however, are making use of other internal states. After all, light hits the retina at the back of the eye, and then the retina turns that light into

electrical signals that are sent to the brain. From that point forward, all other computations are performed on information that is already internal to the cognitive system. In some cases, we can trace a chain of representations all the way back out to states in the world through the set of representing relations that link one relation to the next.

The ability to ground representations in actual perceptual states of the world, though, is complicated by the fact that people are able to represent things that are not present at the moment in the environment. That is, not only can I clearly see my dog in front of me when she is in the room, I can also imagine what my dog looks like when I am at work and she is at home. In this case, I may have some representations in my head that are grounded in others, but at some point, there are states of my cognitive system that are not lawfully related to any existing state of the world at that moment. For that matter, I can believe in all sorts of things like ghosts, unicorns, or the Tooth Fairy that do not exist and never have. Thus, there must be other sources of meaning in mental representations.

A proposal for dealing with the fact that not every concept can be defined solely in terms of a lawful representing relationship to another representation comes from philosophical work on conceptual role semantics (Fodor, 1981; Stich & Warfield, 1994). On this view, a mental representation gets its meaning as a result of the role that it plays within some broader representational system. This proposal acknowledges that the connections among representations may be crucial for their meaning. When I discuss semantic networks in the section on structured representations, we will see how the connections among representational elements influence meaning.

Obviously, not every mental representation in a cognitive system can derive its meaning from its connection to other representational elements (Searle, 1980). As an analogy, imagine that you had an English dictionary, but you didn't actually speak English. You could look up a word in the dictionary, and it would provide you the definition using other words. You could look up those words as well, but that would just give you additional definitions that use other words. Without having the sense of how some of the words relate to things that are not part of language (like knowing that *dog* refers to a certain class of four-legged creatures, or that *above* is a certain kind of relationship in the world), you would not really understand English. Likewise, at least some mental representations need to be grounded.

Types of Representations

There have been many different proposals for mental representations. In this section, I will catalog three key types of representations: mental spaces, featural representations, and structured representations. I discuss each of these representations in a separate section, but before that I want to discuss some general ways that proposals for representations differ from each other (see also Markman & Dietrich, 2000).

First, proposals for representations differ in the presence of discrete symbols. Some representations use continuous spaces. For example, an analog clock represents time using a circle. Every point on that circle has some meaning, though that meaning depends on whether the second hand, minute hand, or hour hand is pointing to that location. In contrast, a digital clock uses symbols to represent time. I will have more to say about symbols when discussing feature representations.

Second, representations differ in whether they include specific connections among the representational elements. Some representations (like the spatial and feature representations I discuss) involve collections of independent representational elements. In contrast, the semantic networks and structured representations discussed later specify relationships among the representational elements.

These two dimensions of difference are generally correlated with the type of representation. In addition, there are dimensions of difference that cross-cut the general proposals for types of representations. For example, representations differ in how enduring the states of the representation are intended to be. Some representations—particularly those that are involved in capturing basic visual information and states of the motor system—are meant to capture moment-by-moment changes in the environment. In contrast, others capture more enduring states.

A final dimension of difference is whether the representation is analog or symbolic. An analog representation is one in which the representing world has the same structure as the represented world. For example, I mentioned that watches represent time using a circle. These watches are analog representations, because both space and time are continuous. Increasing rotational distance in space is used to represent increasing differences in time up to the limit of the time span of the circle.

One advantage of an analog representation is that there are many aspects of the structure of the representing world that can be used to represent

relationships in the represented world without having to define them explicitly. Let us return to the example of using angular distance to represent time, when distances are measured on an absolute scale. If one interval is represented by a 90-degree movement and a second is represented by a 180-degree movement, the first interval moves half the distance of the second. Without having to create any additional relations, we can also assume that the first time interval is half the length of the second.

It is rare to find a representing world that has structure that is similar enough to one in a represented world to warrant creating an analog representation. Thus, most representations are *symbols*: They have an arbitrary relationship between the representing world and the represented world. With symbolic representations, the representing world does not have any structure, and so all of the relationships within the representing world have to be defined explicitly. For example, if we used numbers to represent time, then the representing world of symbols has to define all of the relationships among the symbols for the different numbers to capture relevant relations among numbers in the represented world. For example, the ability to determine that one interval is half the length of another has to be defined into the system, rather than being a part of the structure of the representing world itself.

Finally, it is important to note that an important reason why there are so many different proposals for kinds of knowledge representations is that any choice of a representation makes some processes easy to perform and makes other things difficult to do (see Doumas & Hummel, Chapter 5). There is good reason to believe that the cognitive system uses many different kinds of representations in order to provide systems that are optimized for particular tasks that must be carried out (Dale, Dietrich, & Chemero, 2009; Dove, 2009). Thus, when evaluating proposals about representations, it is probably best to think about what kinds of representations are best suited to a particular process rather than trying to find a way to account for all of cognition with a particular narrow set of representational assumptions. I return to this point at the end of the chapter.

Spatial Representations

DEFINING SPACES

The first type of representation on our tour uses space as the representing world (Gärdenfors, 2000). It might seem strange to think about a mental

representation involving space. We clearly use physical spaces to help us represent things all the time. For example, a map creates a two-dimensional layout that we then look at to get information about the spatial relationships in the world. And I have already discussed analog clocks in which angular distance in space is used to represent time.

Space in the outside world has three dimensions. That means that you can put at most three independent (or *orthogonal*) lines into space. Once you have those three lines set up in your space, you can create an address for every object in the space using the distance along each of those three dimensions. Mathematically, though, a space can have any number of dimensions. In these high-dimensional spaces, points in space have a coordinate for each dimension that determines its location, just as in the familiar three-dimensional space. Objects can then be represented by points or perhaps regions in space.

Once you have located points in space, it is straightforward to measure the distance between them using the formula

$$d(x, y) = \left[\sum_{i=1}^N (x_i - y_i)^r \right]^{1/r} \quad (1)$$

where $d(x,y)$ is the distance between points x and y (each with coordinates for each dimension i , x_i and y_i), N is the number of dimensions, and r is the distance metric, sometimes called the Minkowski metric. When measuring distance in a space, the familiar Euclidean straight-line distance sets $r = 2$. When $r = 1$, then distance is measured using a “city

block” metric in which the distance corresponds to the summed distances along the axes of the space.

It is easy to calculate distance within a space, and so the distance between points becomes an important element in spatial representations. For example, Rips, Shoben, and Smith (1973) tried to map people’s conceptual spaces for simple concepts like birds and animals (see Rips et al., Chapter 11). An example of these spaces is shown in Figure 4.2. The idea here is that pairs of similar birds (like robin and sparrow) are near in space, while pairs of dissimilar birds (like robin and goose) are far away in space. A cognitive process that uses a spatial representation can calculate the distance between concepts when it needs information about the similarity of the items. For example, Rips et al. found that the amount of time that it took to verify sentences like “A robin is a bird” was inversely related to the distance between the points representing the concepts in the space. Consistent with this model, people are faster to verify that the sentence “A robin is a bird” is true than to verify that the sentence “A duck is a bird” is true.

One reason why mental space models of mental representation are appealing is that there are mathematical methods for generating spaces from data about the closeness of items in that space. One technique, called *multidimensional scaling* (MDS), is the one that Rips et al. used in their study (Shepard, 1962; Torgerson, 1965). Multidimensional scaling places points in a space based on information about the distances among those points. For example, if you were to give an MDS algorithm the distances among

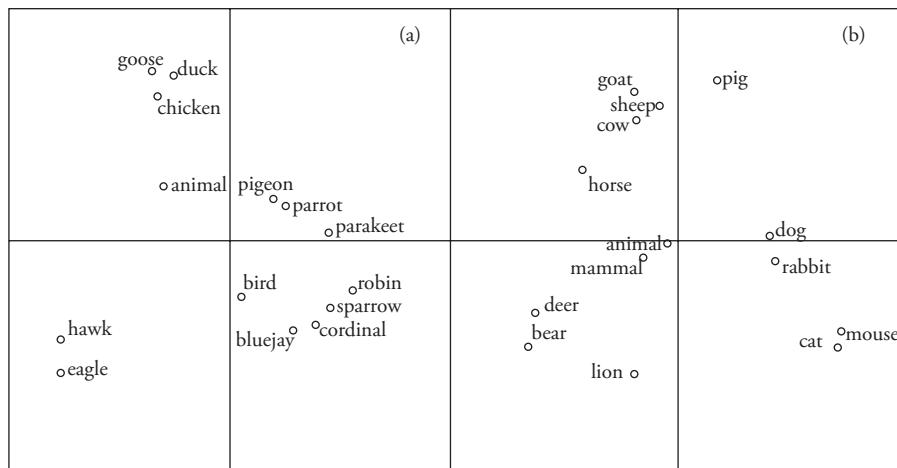


Fig. 4.2 A sample semantic space of a set of concepts representing various birds (taken from Rips, Shoben, & Smith, 1973).

15 European cities, you would get back a map of the locations of those cities in two dimensions.

This same technique can be used for mental distances. For example, Rips et al. had people rate the similarities among the pairs of concepts in Figure 4.2. Similarity ratings can be interpreted as a measure of mental closeness (see Goldstone & Son, Chapter 10). These similarity ratings can be fed into an MDS algorithm. The space that results from this algorithm is the one that best approximates the mental distances it was given. One difficulty with creating spaces like this is that they require a lot of effort from research participants. If there are X items in the space, then there are

$$X(X - 1)/2 \quad (2)$$

distances among those points. That requires a lot of ratings from people, and so in practice it becomes difficult to generate spaces with more than 15 or 20 items in them.

Other techniques have been developed that create very high-dimensional spaces among larger sets of items. A powerful set of techniques uses the co-occurrence relationships among words to develop high-dimensional semantic spaces from corpora of text (Burgess & Lund, 2000; Landauer & Dumais, 1997). These techniques take advantage of the fact that words with similar meanings often appear along with the same kinds of words in sentences. For example, in the sentence “The X jumped up the stairs” the words that can play the role of X are all generally animate beings.

These techniques can process millions of words from on-line databases of sentences from newspapers, magazines, blog entries, and Web sites. As a result, these techniques can create spaces with hundreds or thousands of dimensions that represent the relationships among thousands of words. The interested reader is invited to explore the papers cited in the previous paragraph for details about how these systems operate. For the present purposes, what is important is that after the space is generated, the high-dimensional distances among these concepts in the space are used to represent differences in meanings among the concepts.

STRENGTHS AND WEAKNESSES OF SPATIAL MODELS

There are two key strengths of spatial models of representation. On the practical side, the availability of techniques like MDS and methods for generating

high-dimensional semantic spaces from text provide modelers with a way of creating representations from the data obtained from subjects in studies. For the other types of representations described in this chapter, it is often more difficult to determine the information that ought to end up in the representations generated as parts of explanations of particular cognitive processes.

A second strength of spatial representations is that the mental operations that can be performed on a space are typically quite efficient. For example, there are very simple algorithms for calculating distance among points in a space, and so the processes that operate on spatial representations are quite easy to carry out. Thus, spatial representations are often used when comparisons have to be made among a large number of items.

Despite these strengths, there are also limitations of spatial representations. The core limitation is that the calculation of distance among points provides a measure of the degree of proximity among points. It is also possible to generate vectors that measure a distance and direction between points. However, the underlying dimensions of a semantic space have no obvious meaning. A modeler looking at the representation may ascribe some meaning to that distance and direction, but the system itself has access only to the vector or distance.

People are also able to focus on specific commonalities and differences among pairs of items, and these properties influence people's assessments of similarity. For example, Tversky and his colleagues (Tversky, 1977; Tversky & Gati, 1982) explored the commonalities and differences that affect judgments of similarity. As an example of the kinds of items he used, I'll paraphrase a discussion by William James (1892/1985). When people compare the moon and a ball, they find a moderate degree of similarity, because both are round. When people compare the moon and a lamp, they find a moderate degree of similarity, because both are bright. However, when people compare a ball and a lamp, they find no similarity at all, because they don't share any properties. As Tversky (1977) pointed out, this pattern of similarities is incompatible with a space, because if two pairs of items in a real space are fairly close to each other, then the third pair of points must also be reasonably close. In the mathematical definition of a space, this aspect is called the *triangle inequality*. The observed pattern of similarity judgments by people reflects that they give a lot of weight to specific commonalities among pairs, though different

pairs may focus on different commonalities. Thus, human similarity judgments frequently violate the triangle inequality.

One response to patterns of data that seem inconsistent with spatial representations is to add mechanisms that allow spaces or distances to be modified in response to context (Krumhansl, 1978; Nosofsky, 1986). However, these mechanisms tend to complicate the determination of distance. Because the simplicity of the computations in a space was one of the key strengths of spatial representations, many other theories have opted to use representations with explicit symbols in them to capture cognitive processes that involve a specific focus on particular representation elements. We turn to representations that consist of specific features in the next section.

Feature Representations

The reason why it is difficult to focus on particular commonalities and differences or particular properties in a spatial representation is that spaces are continuous. The core aspect of a feature representation is that it has discrete elements that make up the representation. These discrete elements allow processes to fix reference to particular items within the representation (Dietrich & Markman, 2003).

Typically, feature representations assume that some mental concept is represented by a collection or set of features, each of which corresponds to some property of the items. For example, early models of speech perception assumed that a set of features could be used to distinguish among the various phonemes that make up the sounds of a language (Jakobsen, Fant, & Halle, 1963). On this view, the sound /b/ as in *bog* and the sound /d/ as in *dog* differ by the presence of a feature that marks where in the mouth the speech sound is produced. Linguists identified the particular features that distinguished phonemes by finding pairs of speech sounds that were as similar as possible except for particular features that led them to be distinguished. To return to the example of /b/ and /d/, these phonemes are similar in other properties like engaging the vocal cords (which would distinguish /b/ from /p/).

When this proposal for phoneme representation was being evaluated, a key aspect of the research program for understanding speech perception involved finding some mapping between the features of speech sounds and the speech signal itself (Blumstein & Stevens, 1981). On this view, the speech perception system would identify the features in a particular phoneme from aspects of the

audio signal. A particular phoneme would be recognized when the collection of features associated with it was understood. While this process is straightforward, a weakness of this particular approach to speech perception was that it has proven difficult to isolate particular aspects of the speech signal that reliably indicate particular phonetic features.

Featural models have also been used prominently to study similarity (see Goldstone & Son, Chapter 10). In particular, Tversky's (1977) contrast model assumed that objects could be represented as sets of features. Comparing a pair of representations then required elementary set operations. The intersection of the feature sets of a pair were the commonalities of that pair, while the set differences were the differences of the pair. Tversky proposed that people's judgments of similarity should increase with the size of the set of common features and decrease with the size of the sets of distinctive features. He provided support for this model by having people describe various objects by listing their features. He found that people's judgments of the similarity of various pairs was positively related to the number of features that the lists had in common and negatively related to the number of features that were unique to one of the lists.

Some proposals for feature representations augment the features with other information. For example, it is common to include information about the importance of particular features to a category in a representation. In a classic paper, Smith, Shoben, and Rips (1974) argued that people distinguish between core feature of items and characteristic properties. For example, having feathers is a core characteristic of birds, while singing is typical of birds but is not central for something to be a bird.

They argued that when people classify an object, they look first at all of the properties of the object. If they are uncertain of the category that something belongs to based on its overall similarity, then they focus just on the characteristic properties. That is why people have difficulty classifying objects like dolphins. Dolphins have many features that are characteristic of fish, but they also have features of mammals. Only when people focus on the core characteristics of fish and mammals is it possible to classify a dolphin correctly as a mammal.

Some featural models also include information about the degree of belief in the feature. That is, there are some properties that someone may be certain are true of an object but others for which it is less clear. Some models have proposed that there are

certainty factors that allow a system to keep track of the degree of belief in a particular property (see, e.g., Lenat & Guha, 1990; Shafer, 1996).

While people clearly need information about the degree of belief in a property or the likelihood that the information is true, it is not clear that some kind of marking on features is the best way to handle this kind of information. Often, we want to know more than just how strongly we believe something or how central it is to a category. We want to know *why* a particular fact is central or what it is that causes us to believe it. To support this kind of reasoning, it is useful to have representations that contain explicit connections among the representational elements. The following section discusses types of representations that capture relationships among representational elements.

Structured Representations

Feature representations do a good job of representing the properties of items but a poor job at representing relationships. Consider a variety of relationships that may exist in the world. Poodles are *a kind of* dog. John is *taller than* Mary. Sally *outperformed* Jack. These kinds of relationships are more than just properties of some item. The way that items are connected (or *bound*) to the relationship also matters. Saying that poodles are a kind of dog is true, but saying that dogs are a type of poodle is not.

To capture these relational bindings, structured representations contain mechanisms for creating representations that take arguments that specify the scope of the representation. For example, we can use the notation *kind-of*(?x,?y) to denote that some item?x is a kind of?y. I precede the letter x with a question mark here to denote that it is a variable that can be filled in with some value. Thus, to specify a particular relation, we fill in values for the variables.

Representations of this type are often called *predicates*. Once the values are specified, the representation states a *proposition*, and in a logical system all propositions can be evaluated as true or false, though most proposals for using predicate representations in psychological models do not evaluate the logical form of these predicate structures.

The example *kind-of* (?x,?y) is a *binary* predicate, because it takes two arguments. A predicate that takes one argument, like *red* (?x), is often called an *attribute*, because it is typically used to describe the properties or attributes of objects. This type of

predicate is particularly useful in situations in which there are multiple items in a scene, and it is necessary to bound the scope of the representation.

For example, consider the simple scenes at the top of Figure 4.3. One depicts a circle on top of a square and the other shows a square on top of a circle. In the left-hand scene, one figure is shaded and another is striped. If we just had a collection of features (as shown in the middle of this figure), then it wouldn't be clear which object was striped and which one was shaded. Indeed, the same collection of features could be used to describe both the left- and right-hand scenes, even though they are clearly different.

Because attributes take arguments, though, it is possible to determine the scope of each representational element. The bottom section of Figure 4.3 shows a structured representation of the same pair of scenes drawn as a graph. The relation above (?x,?y) is presented as an oval with lines connecting the relation to its arguments. Likewise, the rounded rectangles are attributes that are connected to the objects they describe. Using this type of representation, it is possible to specify that it is the circle that is striped in the left-hand scene.

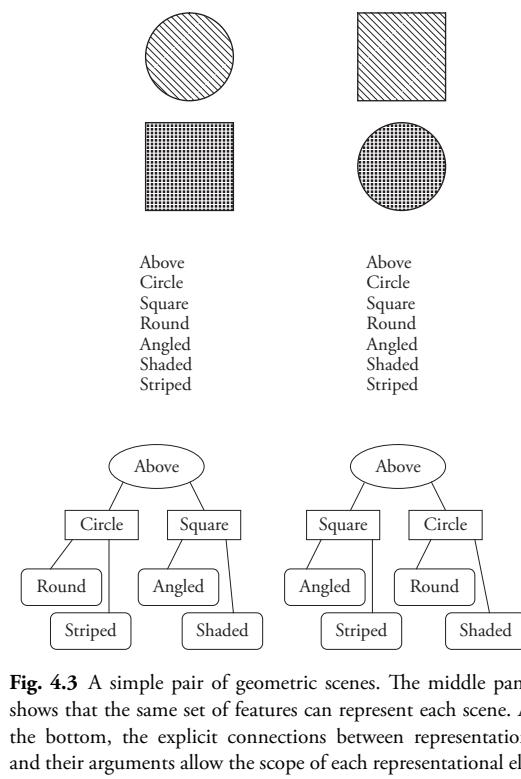


Fig. 4.3 A simple pair of geometric scenes. The middle panel shows that the same set of features can represent each scene. At the bottom, the explicit connections between representations and their arguments allow the scope of each representational element to be defined.

The increase in expressive power that structured representations provide requires that there be processes that are sensitive to this structure. These structure-sensitive processes often require more computational effort than the processes that were specified for spatial and featural representations. In the next two sections, I discuss two different approaches to structured representations that use different processing assumptions to access the connections among representational elements.

SEMANTIC NETWORKS

An early use of structured representations in psychology was semantic networks (Collins & Loftus, 1975; Collins & Quillian, 1972; Quillian, 1968). In a semantic network, objects are represented by nodes in the network. These nodes are connected by links that represent the relations among concepts. The links are directed so that the first argument of a relation points to the second. For example, Figure 4.4 shows a part of a simple semantic network with nodes relating to the concepts vampire and hematologist. This network has a variety of relations in it such as *drinks* (*vampire*, *Blood*) and *studies* (*hematologist*, *Fluid*).

One use of semantic networks was to make simple inferences (Collins & Quillian, 1972). One way to make inferences in a semantic network is to use *marker passing*. In this process, you seek a relationship between a pair of concepts by placing a marker at each of the concepts. The markers are labeled with the concept where they originated. At each step of the process, markers are placed on each node that can be reached from a link that points outward from a node that has a marker on it. When markers

from each concept are placed at the same node, then the path back to the original nodes is traced back, and that specifies the relationship between the concepts.

For example, in the network shown in Figure 4.4, if I wanted to know the relationship between a vampire and a hematologist, I would start by placing markers at the vampire and hematologist nodes. At the first time step, a marker from *vampire* would be placed on the *monster*, *cape*, and *Blood* nodes. A marker from *hematologist* would be placed at the *doctor*, *lab coat*, *water*, and *Blood* nodes. The presence of markers from each of the starting concepts at the *Blood* node would lead to the conclusion that vampires drink blood, while hematologists study blood. The amount of time that it takes to make the inference depends on the number of time steps it takes to find an intersection between paths emerging from each concept.

A second process often used in semantic networks is *spreading activation* (e.g., Anderson, 1983; Collins & Loftus, 1975). Spreading activation theories assume that there are semantic networks consisting of nodes and links, though they do not require the links to be directed. Unlike the marker passing process, activation can spread in both directions along a link.

Each node has some level of activation that determines how accessible that concept is in memory. When a concept appears in the world or in a discourse or in a sentence, then the node for that concept temporarily gets a boost in its activation. That activation then spreads across the links and activates neighboring concepts. Models of this sort have been used to explain priming effects in which processing

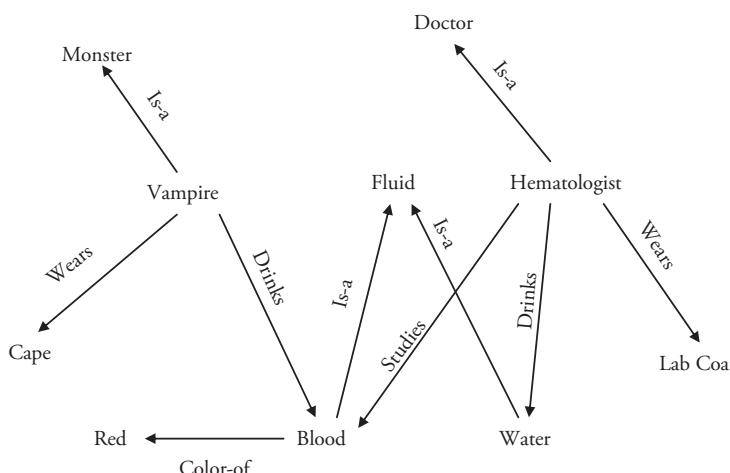


Fig. 4.4 A simple semantic network showing concepts relating to vampires and hematologists (drawn after Markman, 1999).

of one concept speeds processing of related concepts. For example, a classic finding in the priming literature is that seeing a word (e.g., *doctor*) speeds the identification of semantically related words like *nurse* (Meyer & Schvaneveldt, 1971).

Spreading activation theories have been augmented with a number of mechanisms to help them account for more subtle experimental results. For example, links in a network may vary in their strength, which is consistent with the idea that concepts differ in their strength of association. One interesting addition to spreading activation models is the concept of *fan* (Anderson, 1983). The fan of a node in a network is the number of links that leave it, which differs for each node. In some models, the total amount of activation that is spread from one node to others is divided by the number of links, so that nodes with a low fan provide more activation to neighboring nodes than do nodes with high fan. This mechanism captures the regularity that when a node has low fan, then the presence of one concept strongly predicts the presence of the small number of other concepts to which it is connected. In contrast, when a node has a high fan, the presence of one concept doesn't predict the presence of another concept all that strongly.

Semantic network models have been used primarily as models of the relationships among concepts in memory. These processing mechanisms are also shared with interactive activation models that have been used to account for a variety of psychological phenomena (see, e.g., McClelland & Rumelhart, 1981; Read & Marcus-Newhall, 1993; Thagard, 1989, 2000). There is much more that can be done with the relational structure in representations than just passing markers or activation among nodes. I discuss some additional aspects of structured representations in the next section.

STRUCTURED RELATIONAL REPRESENTATIONS

A variety of aspects of psychological functioning seem to rely on people's ability to represent and reason with relations. For example, language use seems to depend crucially on the ability to use relations. Verbs bind together the actors and objects that specify actions (e.g., Gentner, 1975, 1978; Talmy, 1975). Prepositions allow people to talk about spatial relationships (e.g., Landau & Jackendoff, 1993; Regier, 1996; Talmy, 1983). For both verbs and prepositions, the objects that they relate can be specified in sentences that essentially fill in the arguments to these relations.

Related to our ability to talk about complex relations is the ability to reason about the causal structure of the world. Obviously, many verbs focus on how events in the world are caused and prevented (Wolff, 2007; Wolff & Song, 2003). These verbs reflect that people are adept at understanding why events occur. For example, Schank and colleagues (Schank, 1982; Schank & Abelson, 1977) proposed that people form scripts and schemas to represent complex events like going to a restaurant or going to a doctor's office. These knowledge structures contain relationships among the components of an event that suggest the order that things typically happen. They also have causal relations among the events that explain why they are performed (see Buehner & Cheng, Chapter 12).

These causal relations are particularly important for helping people to reason about situations that do not go as expected. For example, when visiting a restaurant, the waiter typically brings you a menu. Thus, you expect to get a menu when you are seated. You also know that you get a menu, because it is necessary to know what the restaurant serves so that you can order food. If you do not get a menu, then you might start to look around for other sources of information about what the restaurant serves like a board posted on a wall.

This range of relations creates an additional burden, because processes have to be developed to make use of this structure. To provide an example of both the great power of these processes as well as their computational burden, I describe work on analogical reasoning. The study of analogical reasoning is a great cognitive science success story, because there is widespread agreement on the general principles underlying analogy, even if there is disagreement about some of the fine details of how analogy is accomplished (Falkenhainer, Forbus, & Gentner, 1989; Gentner, 1983; Holyoak & Thagard, 1989; Hummel & Holyoak, 2003; Keane, 1990; see Holyoak, Chapter 13).

Analogies involve people's ability to find similarities between domains that are not similar on the surface. To find these nonliteral similarities, people seek commonalities in the relations in two domains. For example, a classic example of an analogy is the comparison that the atom is like the solar system. This analogy came out of the Rutherford Model of the atom, which was prominent in the early 20th century. The domain that people know most about is typically called the *base of source* (in this case, the solar system¹), while the domain that is lessunderstood is

called the *target* (in this case, the atom). Atoms are not like solar systems because of the way they look. Atoms are small, and solar systems are large. The nucleus of an atom is not hot like the sun. There are no electrons that support life like planets do. What is similar between these domains is that the electrons of the atom revolve around the nucleus in the same way that the planets revolve around the sun.

Theories of analogy assume that people represent information with structured relational representations. So we could think of the solar system as being represented with some simple relations like

revolve-around (electron, nucleus)
and
greater (mass (nucleus), mass (electron))

which reflect that the electron revolves around the nucleus and that the mass of the nucleus is greater than the mass of the electron. Then, the solar system could be represented by a more elaborate relational system like

cause (greater (mass (sun), mass (planet)),
revolve-around (planet, sun))

These representations might also have a lot more descriptive information about the attributes of the nucleus, electrons, sun, and planets.

The process of analogical mapping seeks parallel relational structures. It does so by first matching up relations that are similar. So the revolve-around ($?x,?y$) relation in the representation of the solar system would be matched to the revolve-around ($?x,?y$) relation in the representation of the atom. Once relations are matched, the arguments of those relations are also placed in correspondence. This constraint on analogy is called *parallel connectivity* (Gentner, 1983), and it is incorporated into many models of analogical reasoning (Falkenhainer et al., 1989; Holyoak & Thagard, 1989; Hummel & Holyoak, 1997; Keane, 1990). So the electron in the atom and the planet in the solar system are matched, because both revolve around something (i.e., they are the first argument in the matched revolve-around ($?x,?y$) relation). Similarly, the nucleus of the atom and the sun in the solar system are matched because both are being revolved around. As many matching relations between domains are found as possible, provided that each object in one domain is matched to at most one object in the other domain (Gentner's, 1983, *one-to-one mapping* constraint).

Analogical mapping processes also allow one domain to be extended based on the comparison

to another. In this process of analogical inference, relations from the base that are consistent with the correspondence between the base and target can be carried over to the target. For example, in the simple representations of the atom and the solar system shown earlier, the match between the domains licenses the inference that the electron revolves around the nucleus because the nucleus is more massive than the electron. This example also demonstrates that inferences drawn from analogies may be plausible, but they need not be true.

One final point to make about analogy is that the structure in representations helps to define which information in a comparison is salient. In particular, some correspondences between domains may match up a series of independent relations that happen to be similar across domains. However, some relations take other relations as arguments. For example, the cause ($?x,?y$) relation in the solar system representation takes two other relations as arguments. Analogies can be based on similarities among entire systems of relations, some of which have embedded relations as arguments. Gentner's (1983) *systematicity* principle suggests that mappings that capture similarities in relational systems are considered to be particularly good analogies (Clement & Gentner, 1991).

This discussion of analogy raises two important general points about structured representations. First, the structure mapping process is complex. Compared to the comparison processes for spatial and featural representations, there are more constraints that have to be specified to describe the structure mapping process. The representation provides few constraints on the nature of the process, and so significant effort has to be put into setting up appropriate processes that act on the representation.

Second, structural alignment is much more computationally intensive than either distance calculations or feature comparison (see Falkenhainer et al., 1989 for an analysis of the computational complexity of structure mapping). Generally speaking, structure-sensitive processes are more computationally intensive than those that can operate on spatial and featural representations. Thus, they are most appropriate for models of processes for which significant time and processing resources are available.

This concludes our brief tour of types of representations. It is not possible to do justice to all of these types of representations in a chapter of this length. I have discussed all of these representations in more detail elsewhere (Markman, 1999, 2002).

Broader Issues About Representation

In the rest of this chapter, I focus on some broader issues and open questions in the area of knowledge representation in the field. I start with an exploration of the way that knowledge is treated in cognitive psychology and how that differs from the treatment of knowledge in other areas of psychology. Then, I discuss some approaches to representation that have focused on the importance of the physical body and of the context when thinking about mental representation. Finally, I discuss a stream of research in the field that has argued that the concept of representation has outlived its usefulness.

Content and Structure

One thing you may have noticed about this entire discussion about knowledge representation is that it has focused on the structure of the knowledge people have without regard to the content of what they know. Spatial representations, featural representations, and structured representations differ in their assumptions about how knowledge is organized. Models of this type have been used in theories that range from vision to memory to reasoning.

It is not logically necessary that the discussion of knowledge representation be organized around the structure of knowledge. Developmental psychology, for example, focuses extensively on the content of children's knowledge (e.g., Carey, 2010). It is quite common to see discussions within developmental psychology that focus on what information children possess and the ages at which they can use that knowledge to solve particular kinds of problems.

Cognitive psychology, however, emerged out of the cognitive revolution. This view of mind is dominated by computation. A computer does not care what it is reasoning about. Right now, my computer is executing the commands to power my word processor so that I can write this sentence. But the computer does not really care what this chapter is about. As long as the data structures within the program function seamlessly with the program, the word processor will do its job. Similarly, cognitive theories have focused on the format of mental data structures with little regard for the content of those structures.

In many areas, though, research in cognitive psychology will need to pay careful attention to content. Because most studies in the field have focused on college undergraduates, research has typically explored people with no particular expertise in the domain in which they are being studied. There is good reason to believe, though, that experts reason

differently from novices in a variety of domains, because of what they know (see e.g., Klein, 2000; Bassok & Novick, Chapter 21). Thus, future progress on core aspects of thinking will require attention both to the content of people's knowledge as well as its structure. There are some exceptions to this generalization, such as the work on pragmatic reasoning schemas (Cheng & Holyoak, 1985) and research that has been done on expertise, but studies incorporating the content of what people know are much more the exception than the rule in the field.

Embodied and Situated Cognition

The computational view of mind that dominated cognitive psychology had another key consequence. There has been a pervasive (if implicit) assumption that the sensory systems provide information to the mind and the mind in turn suggests actions that can be carried out by the motor system. More recently, this assumption has been challenged.

One approach challenges the assumption that cognition is somehow separate from perception and motor control (Barsalou, 1999; Glenberg, 1997; Wilson, 2002). This view, which is often called *embodied cognition*, suggests that understanding cognitive processing requires reorienting the view of mind from the assumption that the mind exists to process information to the assumption that cognition functions to guide action.

An early version of this view came from Gibson's (1986) work on vision. A dominant view of vision in the 1970s was that the goal of vision was to furnish the cognitive system with a veridical representation of the three-dimensional world (Marr, 1982). Gibson argued that the primary goal of vision was to support the goals of an organism. Consequently, vision should be biased toward information that suggests to an animal how it should act. He argued that people tend to see objects not in terms of their raw visual properties, but rather in terms of their *affordances*, that is, with respect to the actions that can be performed on an object. On his view, when seeing a chair, we immediately calculate whether we think we could sit on it, stand on it, or pick it up.

More recent approaches to embodied cognition have provided an important correction to the field, by emphasizing the role of perception and action on cognition. This work demonstrates influences of perception on higher level thinking and also the influence of thinking on perception. For example, Wu and Barsalou (2009) gave people different perceptual contexts and showed how the context

influenced their beliefs about the properties of objects. For example, when people list properties of a lawn, they typically talk about the grass and the greenness of the lawn. When they talk about a “rolled-up lawn,” though, they then talk about the roots, even though the grass in a lawn must also have roots. Findings such as this one suggest that people are able to generate simulations of what something would look like and then to use that information in conceptual tasks.

Conceptual processing may also influence perception. A number of studies, for example, have demonstrated that people’s perception of the slope of a hill is influenced by the perception of how hard it would be to climb up the hill. For example, people wearing a heavy backpack see a hill as steeper than those who are not wearing a heavy backpack (Proffitt, Creem, & Zosh, 2001). However, having a friend with you when wearing a heavy backpack makes the hill look less steep than it would look if you were alone. So your beliefs about the amount of social support you have can affect general perception.

The implication of this work on embodied cognition is that there is no clean separation between the mental representations involved in perception, cognition, and action. Instead, the cognitive system makes use of a variety of types of information. Even tasks that would seem on the surface to involve abstract reasoning often use perceptual and motor representations as well (see Goldin-Meadow & Cook, Chapter 32, for a review of the role of gesture in thought).

A related area of work is called *situated cognition* (Hutchins, 1995; Suchman, 1987). Situated cognition takes as its starting point the recognition that human thinking occurs in particular contexts. The external world in that context also plays a role in people’s thought processes. Humans use a variety of tools to help structure difficult cognitive tasks. For example, because human memory is fallible, we make lists to remind us of information we might forget otherwise. Hutchins (1995) examined the way that navigators aboard large navy ships navigate through tight harbors. He found that—while it is possible to think about the abstract process of navigation—the specific way that navigation teams work is structured by the variety of tools that they have to perform the task.

The embodied and situated cognition approaches have broadened the study of cognition by making clear that a variety of representations both inside

and outside the head influence thinking. The initial work in this area focused on demonstrations of the role of perception, action, and tools on cognitive processes. Ongoing research now focuses on how these types of representations coordinate with the kinds of representations more traditionally studied within cognitive psychology.

Antirepresentationalism and a Call for Pluralism

Despite the centrality of the concept of representation within cognitive psychology, some theorists have argued that the computational approach to mind has outlived its usefulness and thus that cognitive psychology should dispense with the concept of representation (Port & Van Gelder, 1995; Spivey, 2007; Thelen & Smith, 1994). There is an important insight underlying this claim. Much of the research, particularly in the period from the 1960s through the mid-1990s assumed that people did quite a bit of thinking before they performed any actions. For example, a classic model of how intelligent systems could plan actions assumed that people were able to generate complex plans that were then passed to a second system that carried out the plan (Fikes & Nilsson, 1971).

It is clear, however, that people coordinate their actions with the world. While it is often useful to have a general plan for how to carry out a task, that plan needs to be adapted to suit the constraints of the particular environment in which it is carried out. For example, it is fine to set a route to drive from your house to a friend’s house, but if a road is closed for construction, then the route will have to be changed along the way to take into account the actual state of the world.

Research on robotics showed that quite a bit of sophisticated behavior could be generated by having robots that were sensitive primarily to what is going on in the environment at that moment (Beer, 1995; Brooks, 1991; Pfeifer & Scheier, 1999). These systems had representations that would satisfy the minimal criteria for a representation that I described earlier, but they did not keep track of any enduring states of the world that would support complex planning. Nonetheless, these robots could achieve simple goals like traversing a hallway while still avoiding obstacles and adapting to changing circumstances. The initial success of these systems suggested that complex representations might play a minimal role in actual cognitive systems.

As with the work on embodied cognition and situated cognition, these antirepresentationalist

approaches have added valuable new insights to the study of thought. It is clear that paying attention to the way that behavior is carried out on-line is important for helping to constrain what a cognitive system is trying to accomplish. A model that assumes a complete plan can be generated and then passed off to modules that will execute that plan is simply not reasonable as a model of cognition.

At the same time, it is also clear that the hard work of generating theories in cognitive psychology will come from understanding how the variety of approaches to representation are coordinated in systems that carry out a number of complex behaviors. Important work remains to be done on the interfaces among representational systems.

For example, Kuipers (2000) has focused on the question of how to get intelligent agents to navigate through complex spatial environments. His system has a level of representation that responds dynamically to changing environmental conditions. It also has representations that generate general plans for how to get from one location in a large-scale environment to another. In between these levels of representation, the system has representations that generate specific programs for an agent's motor system to suggest how it should go about travelling from one location to the next. This system has been implemented successfully in robots that navigate through environments.

The success of systems such as the one I just described suggests that ultimately cognitive theories need to embrace representational pluralism (Dale et al., 2009; Dove, 2009; Markman & Dietrich, 2000). Each of the systems of representation discussed in this chapter has strengths and weaknesses. Some representations—like spatial representations—are good for carrying information about moment-by-moment states of the world. These representations support simple processes that can act quickly in a changing environment. Other representations excel at storing information for the long term and for creating abstractions across many instances. For example, structured representations support the generation of relational systems that can store the essence of the similarities of domains that seem distant on the surface.

Each form of representation seems particularly well suited to specific kinds of cognitive tasks. There is a tendency when generating theories in science to want to focus on a single approach to explanation. Parsimony is often a desirable characteristic of theories, and so a theory that posits only one form of representation would seem better than a theory

that posits multiple forms of representation. But the mind is likely to make use of a variety of forms of representation. We have a large number of cognitive mechanisms that have evolved over millions of years to accommodate a number of disparate cognitive abilities. There is no reason to think that the representations and processes that are best for understanding low-level vision or basic motor control are also going to be appropriate for handling the kinds of deeply embedded syntactic structures that people encounter when reading academic prose. Thus, rather than trying to create theories that focus on only a single kind of representation, cognitive scientists need to become conversant with many different approaches to representation. The key to developing a successful cognitive model will generally involve finding a set of representations and processes that support robust intelligent behavior.

Future Directions

How can we incorporate more research on the content of people's representations into the study of knowledge representation?

How do representations generated from states of the physical body and aspects of the environment interact with representations of more abstract states?

How can we incorporate and coordinate multiple representations within a single model?

Note

1. Some models of analogy refer to the base as the source (e.g., Holyoak & Koh, 1987).

References

- Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 22, 261–295.
- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4), 577–660.
- Beer, R. D. (1995). Computational and dynamical languages for autonomous agents. In R. F. Port & T. V. Gelder (Eds.), *Mind as motion* (pp. 121–148). Cambridge, MA: The MIT Press.
- Blumstein, S. E., & Stevens, K. N. (1981). Phonetic features and acoustic invariance in speech. *Cognition*, 10, 25–32.
- Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, 47, 139–159.
- Burgess, C., & Lund, K. (2000). The dynamics of meaning in memory. In E. Dietrich & A. B. Markman (Eds.), *Cognitive dynamics* (pp. 117–156). Mahwah, NJ: Erlbaum.
- Carey, S. (2010). *The origin of concepts*. New York: Oxford University Press.
- Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, 17, 391–416.
- Clement, C. A., & Gentner, D. (1991). Systematicity as a selection constraint in analogical mapping. *Cognitive Science*, 15, 89–132.

- Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic priming. *Psychological Review*, 82(6), 407–428.
- Collins, A. M., & Quillian, M. R. (1972). How to make a language user. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 309–351). New York: Academic Press.
- Dale, R., Dietrich, E., & Chemero, A. (2009). Explanatory pluralism in cognitive science. *Cognitive Science*, 33(5), 739–742.
- Dietrich, E., & Markman, A. B. (2003). Discrete thoughts: Why cognition must use discrete representations. *Mind and Language*, 18(1), 95–119.
- Dove, G. (2009). Beyond perceptual symbols: A call for representational pluralism. *Cognition*, 110, 412–431.
- Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41(1), 1–63.
- Fikes, R. E., & Nilsson, N. J. (1971). STRIPS: A new approach to the application of theorem-proving to problem-solving. *Artificial Intelligence*, 2, 189–208.
- Fodor, J. A. (1981). *Representations: Philosophical essays on the foundations of cognitive science*. Cambridge, MA: The MIT Press.
- Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. Cambridge, MA: The MIT Press.
- Gentner, D. (1975). Evidence for the psychological reality of semantic components: The verbs of possession. In D. A. Norman & D. E. Rumelhart (Eds.), *Explorations in cognition* (pp. 211–246). San Francisco, CA: W.H. Freeman.
- Gentner, D. (1978). On relational meaning: The acquisition of verb meaning. *Child Development*, 49, 988–998.
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155–170.
- Gibson, J. J. (1986). *The ecological approach to visual perception*. Hillsdale, NJ: Erlbaum.
- Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences*, 20(1), 1–55.
- Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference learning and discovery*. Cambridge, MA: The MIT Press.
- Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory and Cognition*, 15(4), 332–340.
- Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science*, 13(3), 295–355.
- Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104(3), 427–466.
- Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, 110(2), 220–264.
- Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: The MIT Press.
- Jakobsen, R., Fant, G., & Halle, M. (1963). *Preliminaries to speech analysis*. Cambridge, MA: The MIT Press.
- James, W. (1892/1985). *Psychology: The briefer course*. South Bend, IN: University of Notre Dame Press.
- Keane, M. T. G. (1990). Incremental analogizing: Theory and model. In K. J. Gilhooly, M. T. G. Keane, R. H. Logie, & G. Erdos (Eds.), *Lines of thinking* (Vol. 1, pp. 221–235). London: John Wiley and Sons.
- Klein, G. (2000). *Sources of power*. Cambridge, MA: The MIT Press.
- Krumhansl, C. L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review*, 85(5), 445–463.
- Kuipers, B. (2000). The spatial semantic hierarchy. *Artificial Intelligence*, 119, 191–233.
- Landau, B., & Jackendoff, R. (1993). “What” and “where” in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16(2), 217–266.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato’s problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211–240.
- Lenat, D., & Guha, R. V. (1990). *Building large knowledge-based systems*. San Francisco, CA: Addison Wesley Publishers, Inc.
- Markman, A. B. (1999). *Knowledge representation*. Mahwah, NJ: Erlbaum.
- Markman, A. B. (2002). Knowledge representation. In D. L. Medin & H. Pashler (Eds.), *Stevens handbook of experimental psychology* (3rd ed., Vol. 2, pp. 165–208). New York: Wiley.
- Markman, A. B., & Dietrich, E. (2000). In defense of representation. *Cognitive Psychology*, 40(2), 138–171.
- Marr, D. (1982). *Vision*. New York: W.H. Freeman and Company.
- McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part I, An account of basic findings. *Psychological Review*, 88, 375–407.
- Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90(2), 227–234.
- Nosofsky, R. M. (1986). Attention, similarity and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115(1), 39–57.
- Palmer, S. E. (1978). Fundamental aspects of cognitive representation. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 259–302). Hillsdale, NJ: Erlbaum.
- Pfeifer, R., & Scheier, C. (1999). *Understanding intelligence*. Cambridge, MA: The MIT Press.
- Port, R. F., & Van Gelder, T. (Eds.). (1995). *Mind as motion*. Cambridge, MA: The MIT Press.
- Proffitt, D. R., Creem, S. H., & Zosh, W. D. (2001). Seeing mountains in molehills: Geographical-slat perception. *Psychological Science*, 12(5), 418–423.
- Pylshyn, Z. W. (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences*, 3(1), 111–169.
- Quillian, M. R. (1968). Semantic memory. In M. Minsky (Ed.), *Semantic information processing* (pp. 216–260). Cambridge, MA: The MIT Press.
- Read, S. J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, 65(3), 429–447.
- Regier, T. (1996). *The human semantic potential*. Cambridge, MA: The MIT Press.
- Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, 12, 1–20.
- Schank, R. C. (1982). *Dynamic memory*. New York: Cambridge University Press.
- Schank, R. C., & Abelson, R. (1977). *Scripts, plans, goals and understanding*. Hillsdale, NJ: Erlbaum.

- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424.
- Shafer, G. (1996). *The art of causal conjecture*. Cambridge, MA: The MIT Press.
- Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function, I. *Psychometrika*, 27(2), 125–140.
- Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, 81, 214–241.
- Spivey, M. (2007). *The continuity of mind*. New York: Oxford University Press.
- Stich, S. P., & Warfield, T. A. (Eds.). (1994). *Mental Representation*. Cambridge, MA: Blackwell.
- Suchman, L. A. (1987). *Plans and situated actions: The problem of human-machine communication*. New York: Cambridge University Press.
- Talmy, L. (1975). Semantics and syntax of motion. In J. Kimball (Ed.), *Syntax and semantics* (Vol. 4, pp. 181–238). New York: Academic Press.
- Talmy, L. (1983). How language structures space. In H. L. Pick & L. P. Acredolo (Eds.), *Spatial orientation: Theory, research, and application* (pp. 225–282). New York: Plenum Press.
- Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, 12, 435–502.
- Thagard, P. (2000). *Coherence in thought and action*. Cambridge, MA: The MIT Press.
- Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: The MIT Press.
- Torgerson, W. S. (1965). Multidimensional scaling of similarity. *Psychometrika*, 30(4), 379–393.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4), 327–352.
- Tversky, A., & Gati, I. (1982). Similarity, separability and the triangle inequality. *Psychological Review*, 89(2), 123–154.
- Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin and Review*, 9(4), 625–636.
- Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, 136(1), 82–111.
- Wolff, P., & Song, G. (2003). Models of causation and the semantics of causal verbs. *Cognitive Psychology*, 47, 276–332.
- Wu, L. L., & Barsalou, L. W. (2009). Perceptual simulation in conceptual combination: Evidence from property generation. *Acta Psychologica*, 132, 173–189.

Computational Models of Higher Cognition

Leonidas A. A. Doumas and John E. Hummel

Abstract

Process models of higher cognition come in three basic varieties: traditional symbolic models, traditional connectionist models, and symbolic-connectionist models. This chapter reviews the basic representational and processing assumptions embodied in each of these approaches and considers the strengths and limitations of each.

Key Words: computational models, process models, symbolic models, connectionist models

Models in cognitive science span all three of Marr's (1982) levels (see Holyoak & Morrison, Chapter 1). Normative systems (see Chater & Oaksford, Chapter 2), including the Bayesian framework (see Griffiths et al., Chapter 3), address the level of computational theory; neural models (see Morrison & Knowlton, Chapter 6) are specified at the implementation level; and information-processing models, or process models (perhaps the most common type of model in cognitive science), are specified at the level of representation and algorithm.

Process models of higher cognition come in three basic varieties: traditional symbolic models, traditional connectionist models, and symbolic-connectionist models. This chapter reviews the basic representational and processing assumptions embodied in each of these approaches and considers the strengths and limitations of each.

Symbolic Models

The earliest process models of human cognition were *symbolic* in the traditional sense of the term (e.g., Anderson, 1982; Newell & Simon, 1972), and traditional symbolic modeling continues to be important today (see Gentner & Forbus, 2011;

Taatgen & Anderson, 2008, for reviews). Although the details of specific symbolic models differ, at their core they share the underlying assumption that the mind is a symbol system that is best modeled using symbolic operations on symbolic data structures (see also Fodor & Pylyshyn, 1988).

Symbolic Representations

Any representational system consists minimally of a vocabulary of representational elements (e.g., symbols in a symbolic model or nodes in a neural or Bayesian network) and a set of rules for inferring new statements from existing statements (see Markman, Chapter 4). In order for a representational system to count as symbolic, it must also make it possible to combine its basic representational elements into complex structures capable of expressing an open-ended set of relations (Pierce, 1879, 1903; see also Deacon, 1997).

Traditional symbolic models use various representational formalisms, the most common being propositional notation (or labeled graphs). Propositional notation takes the general form *predicate* (*argument₁*, *argument₂*, ..., *argument_n*), where *predicate* specifies some property or relation, arguments 1...n are the arguments of that predicate,

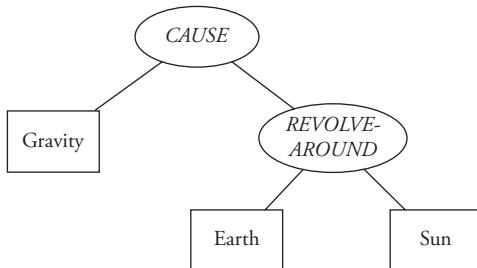


Fig. 5.1 A labeled-graph representation of the higher order relation, *cause*(gravity, *revolve-around*(earth, sun)). Ovals represent relations; rectangles, objects; lines, arcs.

and n is typically three or less. For example, *gave* (John, Mary, book) specifies that John gave Mary a book, and *heavy* (book) specifies that the book is heavy. Labeled graphs specify the same information as propositional notation (i.e., the two systems are isomorphic). In graphical form, nodes represent predicates and their arguments and arcs represent the bindings of arguments to roles of the predicate (see Fig. 5.1). Both formalisms are symbolic in the sense described earlier because they make it possible to form an open-ended (indeed, infinite, as both formalisms permit recursion) set of relational statements with a finite vocabulary of predicates and objects.

Processes

Symbolic representations such as propositions and labeled graphs provide a powerful representational platform that makes many kinds of processes convenient to perform. For example, Forbus, Gentner, and their colleagues (Falkenhainer, Forbus, & Gentner, 1989; Forbus, Gentner, & Law, 1995) have demonstrated that *graph matching*—a process of finding isomorphic substructures in pairs of (potentially very large) systems of labeled graphs—provides an excellent basis for simulating analogical reasoning (i.e., the process of reasoning about a novel *target* domain based on a more familiar *source* domain; see Holyoak, Chapter 13).

Similarly, John Anderson and his colleagues have used propositional notation in their various ACT models (e.g., Anderson, 2007; Anderson & Lebiere, 1998) to very successfully simulate aspects of memory, learning, and inference. The principles embodied in ACT have even been used to develop intelligent tutoring systems that model the learner during the learning process itself, in order to optimize instruction and learning (Anderson, Betts, Ferris, & Fincham, 2010).

In contrast to the graph matching algorithms that Gentner, Forbus, and colleagues have used to model analogical reasoning, the ACT models are based on *production systems*: systems of symbolic rules that operate on propositional knowledge representations to guide action and generate inferences (see Fig. 5.2). Like the graph-matching algorithms of Forbus, Gentner, and colleagues, production systems derive much of their computational power from the fact that they operate on symbolic knowledge structures. For example (see Fig. 5.2), one rule that might be part of a production system is *if* (*larger* (x, y) and *larger* (y, z)) *then* *larger* (x, z). Because this rule is symbolic, it can automatically be applied to any x , y , or z , regardless of their semantic content (e.g., it could infer that a battleship is larger than a submarine from the fact that a battleship is larger than a cruiser and a cruiser is larger than a submarine, and with equal facility make the same inference in the context of a Rottweiler, a housecat, and a mouse). As trivial as this ability might appear, it is a powerful one that cannot not be taken for granted (as we shall see in the context of connectionist models in the next section).

In our previous example, the production rule was completely abstract, defined over empty variables, x , y , and z . But production rules can also be defined over specific arguments (as in, *if* (*see* (me, neighbor's dog), *then run*(me)). Moreover, the rules need not be accurate. For example, a production system used to model the behavior of a young child might include a rule like *if* (*moves*(x), *then* (*has-legs*(x)), reflecting

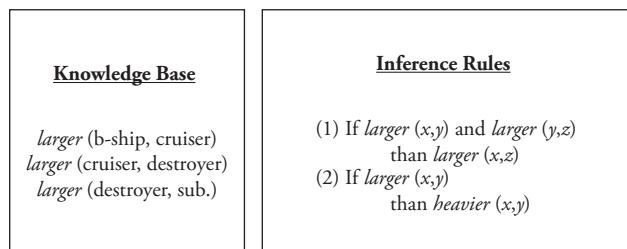


Fig. 5.2 An example of a simple production system.

the child's inaccurate belief that all moving things have legs (e.g., Sheya & Smith, 2006).

Strengths

Symbol systems have been used to successfully model a wide range of cognitive phenomena. The successes of the symbolic approach derive from the flexibility of symbolic representations, in particular the fact that predicates are free to vary in their arguments. As a result, anything that is learned about a predicate (e.g., in the form of a production rule) can, in the limit, be automatically applied to any new argument(s) taken by that predicate.¹ This kind of open-ended generalization—which permits extrapolation beyond the examples from which the rule is learned—is a very powerful inductive mechanism (Holland, Holyoak, Nisbett, & Thagard, 1986; Hummel & Holyoak, 2003). Without it, we would be limited to only those inferences and generalizations that can be found by interpolating over the training examples (Marcus, 2001). That is, without the ability to learn variabilized rules (i.e., rules defined over predicates that are free to vary in their arguments), having learned that relative size is a transitive relation (i.e., the “*if larger...*” rule) in the context of the naval vessels from the example above, we would be at a complete loss to infer *larger* (Rottweiler, mouse) given *larger* (Rottweiler, housecat) and *larger* (housecat, mouse) (see, e.g., Doumas, Hummel, & Sandhofer, 2008; Holyoak & Hummel, 2000; Hummel & Holyoak, 1997, 2003). Penn, Holyoak, and Povinelli (2008; see Penn & Povinelli, Chapter 27) have argued convincingly that this kind of relational generalization is the most important ability distinguishing human cognition from the cognitive abilities of our closest primate cousins.

It is this same capacity for variabilized, relational thinking that allows us to make analogies and to learn and reason from abstract schemas (Holland et al., 1986; Holyoak & Hummel, 2001; Hummel & Holyoak, 2003). Gick and Holyoak (1983) demonstrated that when people draw analogies between similar stories (i.e., specifying which elements of one correspond to which of the other), they may induce generalized schemas describing the shared aspects of those stories. For example, given one story in which Mary is the enemy of Bill and Bill is the enemy of Ted, so Mary regards Ted as her friend, and another story in which Dirk is the enemy of Roger and Roger is the enemy of Jake, so Dirk regards Jake as his friend, one might infer a schema of the general form, if *enemy-of*(person-1, person-2) and *enemy-of*

(person-2, person-3), then *friend-of*(person-3, person-1). If one were then to come across a situation in which Joe was the enemy of Tim, and Tim the enemy of Bill, one could map Joe onto person-1, Tim onto person-2, and Bill onto person-3, and subsequently infer that Bill is a friend of Joe. Notice that this sort of schema-driven inference is formally equivalent to inference based on a production rule in which the antecedent condition (the *if* portion of the rule) is fulfilled and the consequent (the *then* portion of the rule) fires (Anderson & Lebierre, 1998; Newell, 1990).

In the limit, the capacity to represent relational rules (or schemas) makes it possible to represent and reason about universally quantified functions (see Marcus, 2001). For example, consider the rule “ $\forall x,y,z \text{ if } (\text{pass}(x, y), \text{collect}(x, z))$ ” (or, for all x , if x passes y , then x collects z). The rule can apply to any set of x , y , and z , such as a player (x) passing the “GO” square (y) in Monopoly and collecting \$200 (z), or a person (x) passing the border of a country (y) and collecting a passport stamp (z). Without the power of symbolic representations—specifically, without the ability to represent predicates that are free to vary in their arguments—this kind of flexibility would be impossible.

Weaknesses

Despite their successes, a number of criticisms have been leveled against traditional symbolic models of cognition. A very basic one concerns the question of how (and even whether) it is possible to learn symbolic representations. How, for instance, would we learn a relational predicate like *larger*(x, y)? This question is made difficult, in part, by the very property that makes these representations powerful, namely, the fact that they are free to vary in their arguments: Although *larger* (battleship, cruiser) and *larger* (Rottweiler, housecat) express different ideas, they nonetheless express, if not identical, then at least very similar relations. It is our appreciation of this similarity that allows us to grasp that the battleship corresponds to the Rottweiler rather than the housecat.² In turn, the predicate’s ability to remain (at least largely) invariant over changes in its arguments renders it challenging to learn: We are never exposed to disembodied “*larger-ness*”; rather, all those cases where we have had the opportunity to observe an instance of the *larger* relation have presented it in the context of one specific thing being larger than some other specific thing. Given this type of input, how do we ever learn to represent *larger* in a way that is (even

partially) independent of its arguments (Doumas et al., 2008; Kellman, Burke, & Hummel, 1999)? The difficulty of this question—along with the fact that rather than answering it, traditional symbolic modelers have often simply hand-coded their symbolic representations—has led some to wonder whether it is even possible to learn symbolic representations, such as predicates, from nonsymbolic inputs (e.g., early perceptual representations), a concern that has been cited as one of the most significant shortcomings of the symbolic approach (see, e.g., Leech, Mareschal, & Cooper, 2008; Munakata & O'Reilly, 2003; O'Reilly & Busby, 2002; O'Reilly, Busby, & Soto, 2003). Certainly, with tools of traditional symbolic models it is unclear how these representations can be learned in the first place. However, as elaborated later this shortcoming does not mean that symbolic representations cannot be learned at all (e.g., Doumas et al., 2008).

A second limitation of the traditional symbolic approach, as an account of the human cognitive architecture, is that human mental representations have semantic content: They are *about* things, and they somehow naturally capture how those things are similar to and different from one another. By contrast, traditional symbolic approaches to cognition do not (and, worse, *cannot*; Doumas & Hummel, 2004) capture similarity relations among the entities to which symbols refer. For example, the symbols *Rottweiler* and *battleship* fail to specify what these concepts have in common (precious little besides being objects found on Earth) and how they differ (e.g., in animacy, size, and function). Based on the symbols alone, one's best guess about what battleships and Rottweilers have in common is that both require 10 letters to spell (11 in the plural). Moreover, the meanings of various relations seem to apply specifically to individual relational *roles*, rather than to the relation as a whole. As a result, it is easy to appreciate that the agent (i.e., killer) role of *murder* (x, y) is similar to the agent role of *attempted-murder* (x, y), even though the patient roles differ (i.e., the patient is dead in the former case but not the latter); and the patient role of *murder* (x, y) is similar to the patient role of *manslaughter* (x, y), even though the agent roles differ (i.e., the act is intentional in the former case but not the latter).

The semantic properties of human mental representations manifest themselves in countless ways in human cognition, influencing memory retrieval (e.g., Gentner, Ratterman, & Forbus, 1993; Ross, 1987; Wharton, Holyoak, & Lange,

1996), categorization, and reasoning (Bassok, Wu, & Olseth, 1995; Krawczyk, Holyoak, & Hummel, 2005; Kubose, Holyoak, & Hummel, 2002; Ross, 1987). The meanings of relations and their arguments also influence which inferences seem plausible from a given collection of stated facts. For instance, upon learning about a culture in which nephews traditionally give their aunts a gift on a particular day of the year, it is a reasonable conjecture that there may also be a day on which nieces in this culture give their uncles gifts. This inference is based on the *semantic* similarity of aunts to uncles and nieces to nephews, and on the semantics of gift giving, not the syntactic properties of the *give-gift* relation. Given the important role of semantics in the mental representation of relational roles and the objects that fill those roles, an important criterion for a general account for the human cognitive architecture is that the representations on which it is based be able to capture (or at least approximate) that semantic content.

On this point, traditional symbolic models based on varieties of propositional notation and labeled graphs fare poorly. It has been known for a long time that such representational schemes have difficulty capturing shades of meaning and other subtleties associated with semantic content. This limitation was a central focus of the influential critiques of symbolic modeling presented by the connectionists in the mid-1980s (e.g., Rumelhart, McClelland, & The PDP Research Group, 1986). A review of how traditional symbolic models have handled this problem (typically with look-up tables of one sort or another) also reveals that the question of semantics is, in the very least, a thorny inconvenience (see Doumas & Hummel, 2004, for an argument that the problem is more than simply an inconvenience).

Yet a third limitation of traditional symbolic approaches, also cited by the connectionists in the mid-1980s, is that they make no obvious contact with Marr's (1982) third level of analysis, *physical implementation*: It is not at all clear how something like a graph-matcher or a production system would or could be implemented in the brain (but see Anderson, Qin, Jung, & Carter, 2007, for some progress in this direction).

Connectionist Models

Connectionist neural-network models (also referred to as *parallel distributed processing*, or PDP) were motivated in large part by the perceived limitations of the traditional symbolic approach. Neurally inspired

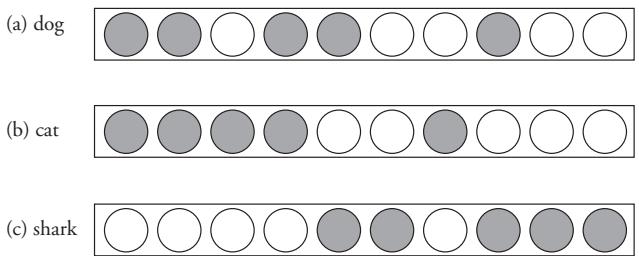


Fig. 5.3 Examples of distributed representations of the concepts, (a) dog, (b) cat, and (c) shark. Note that the same set of units is depicted in each row.

models date back to at least the 1940s (McCulloch & Pitts, 1943; Rosenblatt, 1958). However, their more recent appeal as a general account of the human cognitive architecture—and as a serious alternative to traditional symbolic models—was launched by the work of Rumelhart and colleagues in the mid-1980s (see McClelland et al., 1986; Rumelhart et al., 1986).

Like symbolic models, a variety of connectionist models have been proposed to simulate a wide range of cognitive phenomena; also like symbolic models, the diverse models in the traditional connectionist approach share some basic assumptions about how information is represented and processed.

Representations

Connectionist models consist of collections of simple processors, represented by nodes or *units*, that are connected to form large networks. Units in connectionist networks take activations, typically in the range 0–1. Representations are patterns of activation across *units* in the system. These patterns might correspond to a perception, a thought, a memory, a concept, or any other cognitive state. For example, in the very simple network depicted in Figure 5.3, the pattern of activation on the units depicted in Figure 5.3a might correspond to the concept “dog,” the pattern of activation depicted in Figure 5.3b might correspond to “cat,” and the pattern of activation depicted in Figure 5.3c to “shark.” Other concepts like “animal,” “large,” or “food” would be represented as other patterns.

An important distinction in connectionist models is the distinction between *localist* and *distributed* representations. A localist representation is one in which individual units have meaning (e.g., a unit for “dog”); a distributed representation is one in which the meaning of a concept is carried by a pattern of activation across many separate units (e.g., the concept dog being represented by units for “animal,” “mammal,” “canine,” “domesticated,” etc.). As illustrated by this example, whether a representation is localist or distributed is most often a function of the

relationship between the concept and the representation: With respect to the concept “dog,” units representing “animal,” “mammal,” and so on constitute a distributed representation; but with respect to each of the more general concepts “animal,” “mammal,” and so on, those same units constitute a localist representation. Thus, although the terms “localist” and “distributed” are most commonly used to describe representations (without regard for the entities they represent), they are better thought of as two-place predicates of the form: representation R is localist (or distributed) with respect to concept C (i.e., *distributed* (R, C)). The one exception to this generalization of which we are aware is some vector-symbolic architectures (such as those based on holographic reduced representations; e.g., Plate, 1991), in which the meaning of a unit in any pattern of activation depends entirely on the activity of the other units in that pattern. Such representational schemes are entirely distributed in the sense that no unit in the pattern can be interpreted without reference to the others.

One very appealing aspect of distributed representations (although not those that are entirely distributed) is that they very naturally capture the similarities of different concepts. In our simple network, the concepts of “dog” and “cat” are similar to the extent that their representations overlap. Thus, the network naturally captures the fact that dogs are more similar to cats than to sharks as a natural consequence of its representational scheme. Although there is debate about the utility of distributed and localist codes in connectionist systems (see, e.g., Page, 2000), most connectionist models use some combination of the two (e.g., O’Reilly & Busby, 2002; Rogers & McClelland, 2004).

Processes

Units in a connectionist system are densely interconnected via weighted *connections*. Positive weights act as excitatory connections, so that activity on one unit tends to activate the other, and negative weights act as inhibitory connections, so that activity in one unit tends to reduce activity in the other.

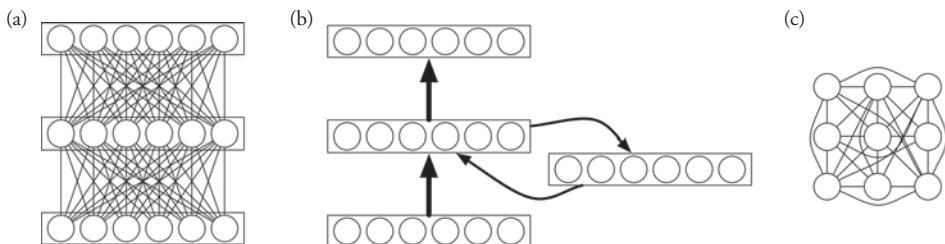


Fig. 5.4 Examples of (a) feed-forward, (b) recurrent, and (c) auto-associative connectionist systems. See text for details.

The architecture of a connectionist network is defined by the manner in which units are connected to one another. The most common architectures are feed-forward (e.g., McClelland & Rumelhart, 1985), recurrent (Elman, 1990), auto-associative (Hopfield, 1982), and hybrids thereof (e.g., O'Reilly, 2006). In a feed-forward architecture, units are arranged in layers such that the units in one layer pass activation (both excitation and inhibition) to the units in the next layer, but units do not typically pass activation to other units in the same layer or in earlier layers (see Fig. 5.4a). Recurrent networks (Fig. 5.4b) are much like feed-forward networks, except that the activation pattern on a “hidden” layer at time t serves as part of the input at time $t + 1$, thereby providing a kind of memory for past states of the network (see, e.g., Elman, 1990). In an auto-associative architecture (Fig. 5.4c) every unit typically has connections to every other unit.

Despite their differences, connectionist networks (at least traditional, nonsymbolic ones as opposed to symbolic ones, as described later) all share the property that the entire currency of computation is activation: Units excite or inhibit one another and over time the state of the network settles into some stable pattern of activation, either on the output units (in the case of feed-forward and recurrent networks) or over the network as a whole (in the case of auto-associative networks). The final or output pattern of activation is interpreted as the network's response to its input.

Strengths

The strengths of the traditional connectionist approach are numerous. Perhaps most obviously, connectionist networks offer a link between cognitive phenomena and their potential neural underpinnings: It is easy to see how the computational principles embodied in a connectionist network could be realized in neurophysiology. In addition, connectionist networks, with their representations based on distributed patterns of activation, provide for a natural kind of automatic generalization. If the representation of,

say, *dog*, consists of one pattern of activation (e.g., on the input units of some network) and the representation of *cat* consists of a similar pattern of activation, then many things learned about dogs will generalize automatically to cats (see McClelland et al., 1986). This kind of automatic generalization accounts for the semantic richness of human mental representations in a way that traditional symbolic representations simply cannot (Doumas & Hummel, 2005).

A third strength of the traditional connectionist approach is that, in contrast to the traditional symbolic approach, it provides a way to seamlessly integrate questions of representation with questions of learning: Connectionist networks are capable of learning their own representations for things, both in the “hidden” layers of feed-forward and recurrent networks, and in classes of unsupervised learning models (which represent a hybrid of feed-forward and auto-associative architectures; see, e.g., Marshall, 1995).

Still a fourth strength of the connectionist approach is that, like biological neural networks, connectionist networks degrade gracefully with damage. As such, they provide a natural platform for simulating the effects of brain damage, normal aging, and even cognitive development (e.g., Colunga & Smith, 2005; Joanisse & Seidenberg, 1999; Li, Lindenberger, & Sikstrom, 2001).

A fifth, less often cited, strength of connectionist models is their mathematical simplicity. In contrast to symbolic models, which are typically complex enough to defy proofs of their computing abilities, connectionist networks are comparatively simple, typically being constrained to using a fixed activation function in a fixed architecture. This makes it possible to prove, for example, that a three-layer (that is, three layers of connections, or a layer of input units, two hidden layers of units, and a layer of output units) feed-forward nonlinear network (where “nonlinear” refers to the activation function of the units in the network) is capable of computing any computable mapping.

Weaknesses

Despite their strengths, traditional connectionist models have some important limitations as models of higher cognition in humans. Most notably, they lack symbolic competence. Many authors have written extensively on this limitation (e.g., Doumas et al., 2008; Hummel, 2000; Hummel & Biederman, 1992; Hummel & Holyoak, 1997, 2003; Marcus, 1998, 2001; von der Malsburg, 1981, 1999), but its importance remains underappreciated.

Although a connectionist network can compute any computable mapping, it effectively must be trained on each individually. Although a properly trained NN can interpolate between learned input-output mapping, it cannot extrapolate to mappings that lie outside of its training set (e.g., Marcus, 1998; St. John, 1992; St. John & McClelland, 1990). Symbolic systems by contrast including humans extrapolate easily (e.g., as in the case of human relational reasoning; Hummel & Holyoak, 2003).

The fundamental reason why connectionist models fail to achieve symbolic competence is that traditional connectionist representations simply do not have enough degrees of freedom—they is, they are formally too weak (Hummel et al., 2004). A symbolic representation, such as propositional notation, has two degrees of freedom with which to express information. The first is the choice of which symbols to use: If a modeler wishes to represent that John loves Mary, she would use the symbols *loves*, John and Mary; to represent that John hates Mary, she would use *hates*, John and Mary. The corresponding degree of freedom in a connectionist representation is activation: To represent that John loves Mary, a connectionist network might activate units (or, in the distributed case, collections of units) representing *loves*, John and Mary.

The second degree of freedom in a symbolic representation specifies the bindings of arguments to the roles of the relation: To represent that John loves Mary, it is traditional to place John in the first slot inside the parentheses following the relation and Mary in the second slot, forming *loves* (John, Mary); to specify that Mary loves John, the modeler would place the *very same symbols* in the opposite slots to form *loves* (Mary, John). (An explanation for the italics appears two paragraphs later.) There is no analogous second degree of freedom in a traditional connectionist representation. Activating units or patterns for *loves*, John and Mary, could equally represent “John loves Mary” or “Mary loves John” (or, in the case of a distributed representation, a statement

about a narcissistic hermaphrodite; Hummel & Holyoak, 1997; von der Malsburg, 1981), so it is impossible to tell which is the intended meaning of the representation. The reason is that this representation fails to specify the bindings of the arguments to the roles of the relation. Unlike the symbolic representation, the traditional connectionist representation lacks a degree of freedom with which to specify this information.

The most common response to this problem is to use varieties of *conjunctive coding* to carry binding information. Under conjunctive coding, units represent, not objects or relational roles, but *conjunctions* of objects in relational roles. For example, rather than representing *loves*, John and Mary, the units in a conjunctive code might represent conjunctions such as *John+lover*, *Mary+beloved*, *Mary+lover*, and *John+beloved*. Now, to represent that John loves Mary, the network activates *John+lover* and *Mary+beloved*; to represent that Mary loves John, it activates *Mary+beloved* and *John+lover*. Specific varieties of conjunctive coding include tensor products (Halford, Wilson, & Phillips, 1998; Smolensky, 1990), holographic reduced representations (Plate, 1991), and spatter codes (Kanerva, 1998). These approaches vary in their particulars, but they all share the property that units represent specific conjunctions of roles and arguments rather than representing the roles or arguments individually.

Models based on conjunctive coding have been applied with varying degrees of success in various domains. And indeed, some kind of conjunctive coding is certainly necessary for encoding bindings in long-term memory (Doumas et al., 2008; Hummel & Biederman, 1992; Hummel & Holyoak, 1997, 2003; Shastri, 2002). But as a general, or the only, solution to the binding problem in connectionist or neural networks, conjunctive coding is sharply limited. Recall the italicized *very same symbols* wording from two paragraphs earlier. It is the fact that the symbols in *loves* (John, Mary) are the same as those in *loves* (Mary, John) that allows you to know what these statements have in common: Both are about loving, John and Mary. The same is true for the statements *jonks* (grummond, steplock) and *jonks* (steplock, grummond): Although these statements are largely meaningless, because they use and reuse the same symbols, in the very least we know that, whatever “jonking” is, the grummond is doing it (or else stands in that relation) to the steplock in one case and the steplock is doing it (or stands in that relation) to the grummond in the other.

This ability to preserve the identity of symbols across different relational structures is a characteristic of symbolic representations, made possible by the fact that they have two degrees of freedom, that is absolutely essential for symbolic thought—such as reasoning based on analogies, schemas, or rules (Holyoak & Hummel, 2002; Hummel & Holyoak, 1997, 2003; Penn et al., 2008). It is for this reason that there are no successful traditional connectionist models of analogy (see Leech, Mareschal, & Cooper, 2008, for an attempt, the shortcomings of which illustrate the need for that second degree of freedom).

A number of connectionist models have been developed that appear to solve the problem of relational processing. We argue that these models in fact fail to simulate human relational reasoning and instead simulate a proxy to relational reasoning.

A connectionist model developed by O'Reilly and Busby (2002) illustrates what can (and cannot) be achieved without the second degree of representational freedom enjoyed by symbolic systems. O'Reilly and Busby's model is designed to answer questions about the spatial relations among objects. The model consists of "input/output" units representing (a) object features at each of 4x4 locations in the visual field, (b) the locations of those objects,

(c) the objects' identities (independent of location), and (d) relations among the objects (e.g., above/below). There are also "query units" associated with the input/output units (used for querying the model after training). These units are connected to (and communicate via) a set of "hidden" units, forming a large auto-associative network. The model is trained by pairing patterns of activation across the input/output units and learning connections between those units and the hidden units using the Leabra learning algorithm (O'Reilly, 1996). After training, the model can be presented with objects in various locations and be asked questions about them. For example, in order to ask, "What is at location 3,2?" the location query unit would be activated, along with the unit for location 3,2. No units in the "object" or "relation" arrays would be activated. The model's task would be to activate the object units corresponding to the distributed representation of whatever object resides at location 3,2. Or if asked, "What is above?" the "above" query unit would be activated, and the model's task would be to activate the representation of the object that is above the other, i.e., that in location 2,2 (see Fig. 5.5).

O'Reilly and Busby (2002) trained the model on various subsets (from 1.3% to 25%) of the input/

Location (1,1)	Location (1,2)	Location (1,3)	Location (1,4)
Location (2,1)	Location (2,2) 	Location (2,3)	Location (2,4)
Location (3,1)	Location (3,2) 	Location (3,3)	Location (3,4)
Location (4,1)	Location (4,2)	Location (4,3)	Location (4,4)

Fig. 5.5 Example of the task simulated by the O'Reilly and Busby (2002) model. A circle is presented in location 2,2, and a square in location 3,2. The model might be asked, "What is above location, 3,2?" (which should activate location 2,3).

query conjunctions it is capable of representing, and tested it for its ability to generalize to the untrained input/query conjunctions. After training on 25% of its input space, the model successfully generalized to roughly 95% of the untrained inputs. Based on this performance, O'Reilly and Busby concluded that "... rich distributed representations containing coarse-coded conjunctive encodings can effectively perform binding" (p. 6).

However, this claim, and the simulations on which it is based, deeply underestimate the power of relational perception, thought, and generalization. Being able to answer "What is above?" is not the same as being able to represent " x is above y ," for all x and all y , and being able to draw inferences based on the latter—for example, if x is above (and on) y , then x could potentially fall off of y . Far from this, the O'Reilly and Busby (2002) model can simply say "what is above": it can answer " x is above" (provided the features of x were part of its training regime), but it cannot say *what x is above*: It lacks even the basic capacity to specify the second argument of the *above* relation.

What makes human relational perception and thought powerful (and difficult to model) is not our ability to answer simple questions of the form "What is above?" (a question the O'Reilly & Busby, 2002, model answers based strictly on associative learning), but our ability to represent relations such as *above* (x, y), explicitly, to bind arguments to the roles of those relations, and to use that knowledge to make inferences about x and y for all x and y (not just those whose features have appeared in our training space). The O'Reilly and Busby model, by contrast, can only answer questions whose answers it has already been taught. Like other traditional connectionist models, the O'Reilly and Busby model can only generalize to new patterns by interpolating among patterns on which the model has been explicitly trained (e.g., patterns that are linear combinations of the examples on which the model has already been trained). And it cannot even represent, much less answer, questions of the form, "If x is in top of y , can x fall off of y or can y fall off of x ?", a question even a 3-year-old child would look at you askance for asking her in the first place.

This limitation is not restricted to the O'Reilly and Busby (2002) model, but is true of all models that have tried to tackle symbolic processing using strictly associationist tools (e.g., Rogers & McClelland, 2004, 2008; St. John & McClelland, 1990). As noble as these efforts are, their tools simply are not up to the task.

Symbolic-Connectionist Models

In response to the complementary strengths and weaknesses of the traditional symbolic and connectionist approaches, some researchers have attempted to implement symbolic structures within connectionist architectures with distributed representations. In principle, achieving symbolic competence in a connectionist system should not be difficult: All that is needed is some basis for representing role-filler bindings in a way that allows the representation of the roles and fillers to remain invariant (i.e., to be "reused" as they are in symbolic systems) across different bindings. Proposed solutions to this problem come in two general forms.

Models Based on Vector Multiplication

One approach to implementing symbolic structure in connectionist systems is to use tensor products (Smolensky, 1990; see also Halford et al., 1994; Halford et al., 1998)—or their variants, such as holographic reduced representations (HRRs; Plate, 1991), spatter codes (Kanerva, 1998), or circular convolutions (Metcalfe, 1990)—to represent role-filler bindings. A tensor product is an outer product of two or more vectors (i.e., a matrix) that is treated as an activation vector rather than a matrix (Smolensky, 1990). For example, to bind a one-place predicate, \mathbf{r} (such as *eats* (x) or *runs* (x)), to its argument, \mathbf{f} , the tensor \mathbf{rf} is formed by multiplying the i th element of vector \mathbf{r} , representing the role, by the j th element of \mathbf{f} , representing the filler (for all combinations of i and j):

$$\mathbf{rf}_{ij} = \mathbf{r}_i \mathbf{f}_j. \quad (1)$$

There are two ways to bind multiplace relations to their arguments using tensor products. One is to define tensors of progressively higher rank (where the rank of a tensor is the number of vectors that come together to define it; see, e.g., Halford et al., 1994). For example, a two-place relation (such as *loves* (x, y) or *larger-than* (x, y)) could be represented by the rank three tensor \mathbf{rfg} :

$$\mathbf{rfg}_{ijk} = \mathbf{r}_i \mathbf{f}_j \mathbf{g}_k \quad (2)$$

where \mathbf{r} represents the relation (e.g., *loves* in *loves* (x, y)), \mathbf{f} represents the argument bound to the first role of the relation (x), and \mathbf{g} represents the argument bound to the second (y). An alternative approach

is to designate separate tensors for each role-filler binding and represent the complete proposition as their sum (e.g., Tesar & Smolensky, 1994). For example:

$$\mathbf{rf} = \mathbf{rf}^1 + \mathbf{rf}^2, \quad (3)$$

where \mathbf{rf}^1 is a tensor representing the binding of the first role to its argument and \mathbf{rf}^2 represents the binding of the second role to the second argument.

The deep problem with all these approaches is that, because tensors are a variety of conjunctive coding, they violate role-filler independence. A tensor is a *product* of two or more vectors (see Eqs. 1–3), so the similarity of two tensors—and hence the ability to generalize something learned about one of them to the other—scales with the *product* of the similarities of the simple vectors from which the tensors were created. For example, if vector similarity is defined in terms of the inner (“dot”) product, then:

$$\mathbf{rf}_1 \cdot \mathbf{rf}_2 = (\mathbf{r}_1 \cdot \mathbf{r}_2)(\mathbf{f}_1 \cdot \mathbf{f}_2), \quad (4)$$

where \mathbf{rf}_1 and \mathbf{rf}_2 are tensors made from \mathbf{r}_1 and \mathbf{f}_1 and from \mathbf{r}_2 and \mathbf{f}_2 , respectively, and \cdot indicates the dot product. Similarly, if vector similarity is defined in terms of the cosine of the angle between two vectors, then:

$$\cos(\mathbf{rf}_1, \mathbf{rf}_2) = \cos(\mathbf{r}_1, \mathbf{r}_2)\cos(\mathbf{f}_1, \mathbf{f}_2). \quad (5)$$

Under both these definitions of vector similarity, two tensors will be similar to one another only to the extent that both their roles *and* fillers are similar: Identical roles, bound to completely different fillers, result in completely different tensor products, precluding any generalization from one to the other. This property is true, not only of tensor products, but of any binding scheme based on tensors (such as HRRs; Plate, 1991).

This property of tensors reflects the fact that they are the result of vector multiplication: Since role and filler vectors are multiplied by one another to form role-filler bindings, the similarities of those vectors are multiplied to determine the similarity of the resulting tensors. This observation suggests that one way to avoid role-filler interaction in role-filler binding similarity is to perform role-filler binding by role-filler addition rather than multiplication.

Models Based on Vector Addition

A second approach to implementing symbolic structure in connectionist networks is to bind roles to fillers by vector addition, which can be implemented as synchrony of neural firing (see Hummel & Holyoak, 1997, 2003). The basic idea is that units representing a relational role are added to (i.e., fire in synchrony with) the units representing the argument filling that role; units representing separate role-filler bindings fire out of synchrony with one another. A related proposal is to represent role-filler bindings by systematic *asynchrony* of firing, such that units representing a relational role fire, for example, just before the units representing its filler, and separate bindings fire in more distant temporal relations (see Doumas & Hummel, 2005; Doumas et al., 2008; Love, 1999). In either case, the binding is represented as a kind of vector addition because a given vector always represents a given role or object, regardless of the object or role to which it happens to be bound, and the bindings of roles to their fillers are represented by operations (synchrony or systematic asynchrony of firing) that put roles together with their fillers in an additive rather than multiplicative fashion.

However, synchrony of firing cannot be the whole story concerning the neural basis for binding, because temporal patterns of neural activity are necessarily transient. At a minimum, conjunctive codes are necessary for the purposes of storing bindings in LTM, and for forming localist tokens of roles, objects, role-filler bindings, and complete propositions (Hummel & Holyoak, 1997, 2003). A complete account of the human cognitive architecture must incorporate both dynamic binding (for independent representation of roles bound to fillers in WM) and conjunctive coding (for LTM storage and token formation), and specify how these coding systems are related.

One example of a system that binds via vector addition is a model of analogical reasoning called Learning and Inference by Schemas and Analogies, or LISA (see also the SHRUTI model; Shastri & Ajjanagadde, 1993). In LISA, relational structures are represented by a hierarchy of distributed and localist codes (see Fig. 5.6). At the bottom, “semantic” units represent the features of objects and roles in a distributed fashion. At the next level, these distributed representations are connected to localist predicate-and-object units (POs) representing individual predicates (or relational roles) and objects. Localist role-binding units (RBs) link object and

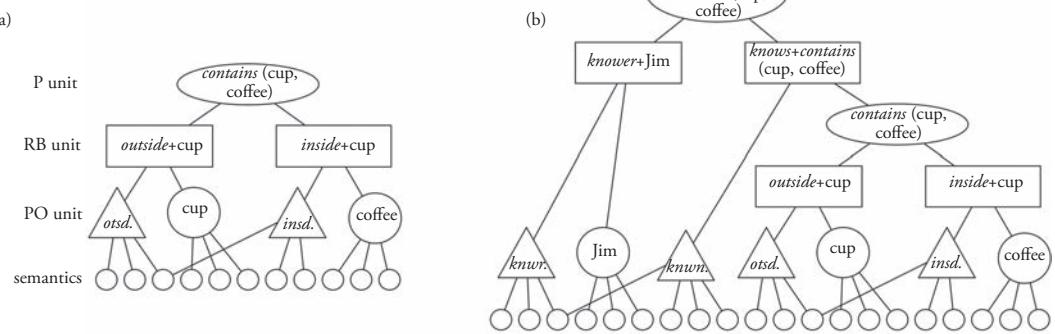


Fig. 5.6 Representation of propositions in LISA (Learning and Inference by Schemas and Analogies). Objects and relational roles are represented both as patterns of activation distributed over units representing semantic features (*semantic units*; small circles) and as localist (PO) units representing tokens of objects (large circles) and relational roles (triangles). Roles are conjunctively bound to fillers by localist role-binding (RB) units (rectangles), and role-filler bindings are conjunctively bound into complete propositions by localist P units (ovals). (a) Representation of *contains* (cup, coffee). (b) Representation of *knows* (Jim, *contains* (cup, coffee)). When one proposition takes another as an argument, the lower (argument) proposition serves in the place of an object unit under the appropriate RB of the higher level P unit (in this case, connecting *contains* (cup, coffee) to the RB representing what is known). In the figure we represent these localist units (i.e., POs, RBs, and Ps) with different shapes for the purposes of clarity. Importantly, these units are not different kinds of units (i.e., they do not work differently). Rather, they are simply units in different layers of the network. We use different names for the units in each layer (and different shapes in the figures) only to make them easier to distinguish.

relational role units into specific role-filler bindings. At the top of the hierarchy, localist proposition units (Ps) link RBs into whole relational propositions.

To represent the proposition *contains* (cup, coffee), PO units (triangles and large circles in Fig. 5.5) representing the relational roles *outside* and *inside*, and the fillers cup and coffee, are connected to semantic units coding their semantic features. RB units (rectangles) then conjunctively code the connection between roles and their fillers (one RB connects cup to *outside*, and one connects coffee to *inside*). At the top of the hierarchy, P units (oval) link sets of RBs into whole relational propositions. A P unit conjunctively codes the connection between the RBs representing *outside+cup* and the RB representing *inside+coffee*, thus encoding the relational proposition *contains* (coffee, cup). Each level of the representational hierarchy serves an important purpose. The semantic units capture the semantically rich (i.e., distributed) nature of human mental representations. The three layers of localist units make it possible to treat each level of the hierarchy as an independent entity for the purposes of mapping and inference (Hummel & Holyoak, 1997, 2003).

When a proposition enters working memory, role-filler bindings must be represented dynamically on the units that maintain role-filler independence (i.e., POs and semantic units; see Hummel & Holyoak, 1997). In models using synchrony of

firing as a binding tag, roles are dynamically bound to their fillers by synchrony of firing (see earlier). In models using systematic asynchrony of firing as a binding tag, roles and their fillers fire in direct sequence. Binding information is carried in the proximity of firing (e.g., with roles firing directly before their fillers).³

An important consequence of the approach is that it allows a solution to the problem of how the structured (i.e., symbolic) representations that underlie symbolic systems may be learned in the first place (as noted earlier, one of the most oft-levied criticisms of the symbolic account of cognition). DORA (*Discovery of Relations by Analogy*; Doumas et al., 2008) is an account of how children and adults learn novel relational concepts from examples and subsequently use those representations in the service of understanding and reasoning about the world. Unlike other models that represent and employ structured representations, DORA learns structured relational representations from unstructured (i.e., nonsymbolic and nonrelational) examples. DORA starts with unstructured representations of objects as simple vectors of features. When DORA compares two or more of these objects, it learns explicit representations of any properties they share. Because DORA can use time as a binding tag (see earlier), the resulting representations are effectively single-place predicates (represented in a distributed fashion) that can

be bound to novel arguments. DORA then combines sets of these predicates to form representations of complete multiplace relations (where each of the combined predicates serves as a role of the new relation). For example, DORA will learn predicates like *inside(x)* and *outside(y)* when it compares examples of different objects inside and outside one another (e.g., children outside a house and coffee inside a cup). DORA will then combine the *inside(x)* and *outside(y)* predicates to form the multiplace relation *contains(x, y)*. Importantly, this is precisely the learning trajectory that children follow when they learn relations (e.g., Smith, 1984). Thus, DORA provides an account of how structured representations can be learned from examples, and so it addresses one of the major criticisms levied at structured models of cognition.⁴

Based on a small set of basic principles (notably comparison-based learning and constructing multiplace relations from sets of single-place predicates), DORA accounts for many phenomena surrounding the development of relational thinking, the development of the shape bias, and the effect of labeling on relational category learning (e.g., Doumas & Hummel, 2010; Doumas et al., 2008; Sandhofer & Doumas, 2008; Son, Doumas, & Goldstone, 2010).

Strengths

Models based on symbolic-connectionist enjoy a range of strengths. Because they are structured (i.e., they have two representational degrees of freedom in that they solve the binding problem), symbolic-connectionist models support the flexible and powerful generalizations of traditional symbol systems. Predicates are free to vary in their arguments, and thus what is learned about a predicate in one context potentially generalizes to other contexts and other arguments (see later discussion). So, like symbolic models, symbolic-connectionist models can extrapolate beyond training examples. The structure sensitivity of symbolic-connectionist models has allowed them to account for a wide range of phenomena in higher order cognition, including analogy making and retrieval from long-term memory (Hummel & Holyoak, 1997), relational generalization and schema induction (Hummel & Holyoak, 2003), learning relational concepts (Doumas et al., 2008), object recognition and the role of attention in shape perception (Hummel, 2001; Hummel & Biederman, 1992), motivation (Sun, 2009), speech production (Chang, Dell, & Bock, 2006), explicit skill learning (Sun, Slusarz, & Terry, 2005), and

creative problem solving (Helie & Sun, 2010). There are also models of this sort that are able to learn structured (i.e., symbolic) representations from unstructured examples (Doumas et al., 2008).

Just as their structure sensitivity affords them the advantages of symbolic competence, so the distributed representations in symbolic-connectionist models afford them many of the strengths of traditional connectionist models. Like traditional connectionist models (as discussed earlier), symbolic-connectionist models offer a link between algorithmic models and possible neural underpinnings, allow natural automatic generalization, integrate representation and learning, and degrade gracefully with damage. Symbolic-connectionist models have been used to account for implicit skill learning (Sun et al., 2005), categorization, reflexive inference (Shastri & Ajjanagadde, 1993), and cognitive deficits due to brain damage (Morrison et al., 2004) and normal aging (Viskontas et al., 2004).

Weaknesses

Symbolic-connectionist models also have their share of weaknesses. One of the more serious shortcomings of symbolic-connectionist models is that they are weaker than purely symbolic models. For example, whereas symbolic models can deal with executive function and goal setting, representing and responding to negation and quantification (e.g., universal—for all—and existential—there exists), these phenomena have yet to be successfully modeled within symbolic-connectionist architectures. It remains unclear how (or even whether) symbolic-connectionist models will be able successfully account for them.

A second shortcoming of symbolic-connectionist models is that their representational assumptions (and thus their resulting representations) are more complex than traditional connectionist models. Although symbolic-connectionist models are connectionist in spirit, in that they are built out of collections of simple processing units that are highly interconnected, they include additional processing assumptions (such as sensitivity to temporal patterns of firing) that allow them to solve the binding problem (i.e., that provide them an additional informational degree of freedom comparable to that available to symbolic models). But whereas it might appear that symbolic-connectionist models make more nativist assumptions than traditional connectionist models (in that the former assumes an additional informational degree of freedom by which to

carry binding information), we would argue that this appearance is misleading. Symbolic-connectionist models make assumptions about learning rules and activation functions similar to those made by traditional connectionist models. The additional assumption that symbolic-connectionist models (at least those that use time to carry binding information) must make beyond traditional connectionist models is that units are organized in layers and are sensitive to what layers they are in.

Conclusions and Future Directions

An adequate account of human mental representations—and the human cognitive architecture more broadly—must account both for our ability to represent the semantic content of relational roles and their fillers, and for our ability to bind roles to their fillers dynamically without altering the representation of either. Traditional symbolic approaches to cognition fail to specify the semantic content of roles and their fillers—a failing that, as noted by the connectionists in the 1980s, renders them too inflexible to serve as an adequate account of human mental representations. Traditional distributed connectionist approaches have the opposite problem: They succeed in capturing the semantic content of the entities they represent but fail to provide any basis for binding those entities together into symbolic (i.e., relational) structures. This failure renders them incapable of relational generalization, which appears to be required for such human abilities as assessing similarity based on alignment (Goldstone & Son, Chapter 10), analogical reasoning (Holyoak, Chapter 13), and learning problem schemas (Bassok & Novick, Chapter 21).

By contrast, symbolic-connectionist models combine the strengths of both the symbolic and connectionist approaches. These models have the potential to produce representations that are neurally plausible, semantically rich, flexible, and meaningfully symbolic. Armed with these representations, the symbolic-connectionist approach may be able to provide a powerful foundation for understanding human cognition. Nonetheless, models to date have their limitations. Those models based on vector multiplication (using tensor products and other forms of conjunctive coding as the sole basis for role-filler binding) fail to capture the natural pattern of similarities among propositions. There remain many important aspects of human cognition that symbolic-connectionist models have not yet addressed, including planning, quantification, negation, and

other aspects of language use. It remains to be seen whether these are simply questions or represent fundamental limitations of the approach.

Notes

1. We say “in the limit” here because this statement assumes that the predicate is defined over variables open enough to take any kind of arguments, such as the x , y , and z in the larger... production rule. By contrast, a production rule defined over specific objects, such as the “me and my neighbor’s dog” rule, would not necessarily be expected to generalize automatically to all new arguments (although used as part of an analogy, it might be expected to generalize very substantially; see, e.g., Holyoak & Thagard, 1995; Hummel & Holyoak, 2003).

2. Note that this correspondence is not based on anything as trivial as the arguments’ locations within the parentheses: The same correspondence is also suggested by the propositions larger (battleship, cruiser) and smaller (housecat, Rottweiler)—a correspondence that also suggests that the semantics of the larger and smaller relations reside, not in the relations as holistic entities, but in their individual roles (Doumas & Hummel, 2005; Doumas, Hummel, & Sandhofer, 2008; Hummel & Holyoak, 1997, 2003).

3. Asynchrony-based binding allows role and filler to be coded by the same pool of semantic units, which allows a system to learn representations of relations from representations of objects (Doumas et al., 2008).

4. As noted by Holyoak (Chapter 13), DORA does not provide an account of what the features of relations are. Rather, it provides an account of how those features (assuming they exist) can be extracted from unstructured representations of objects and represented as explicit structures that can take arguments (i.e., as predicates).

References

- Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, 89, 369–403.
- Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.
- Anderson, J. R., Betts, S. A., Ferris, J. L., & Fincham, J. M. (2010). Neural imaging to track mental states while using an intelligent tutoring system. *Proceedings of the National Academy of Sciences USA*, 107(15), 7018–7023.
- Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.
- Anderson, J. R., Qin, Y., Jung, K-W., & Carter, C. S. (2007). Information-processing modules and their relative modality specificity. *Cognitive Psychology*, 54(3), 185–217.
- Bassok, M., Wu, L., & Oseeth, K. L. (1995). Judging a book by its cover: Interpretive effects of content on problem-solving transfer. *Memory and Cognition*, 23, 354–367.
- Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, 113, 234–272.
- Colunga, E., & Smith, L. B. (2005). From the lexicon to expectations about kinds: A role for associative learning. *Psychological Review*, 112, 347–382.
- Deacon, T. W. (1997). *The symbolic species: The co-evolution of language and the brain*. New York: W. Norton and Company.
- Doumas, L. A. A., & Hummel, J. E. (2004). A fundamental limitation of symbol-argument- argument notation as a

- model of human relational representations. In *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society* (pp. 327–332). Mahwah NJ: Erlbaum.
- Doumas, L. A. A., & Hummel, J. E. (2005). A symbolic-connectionist model of relation discovery. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society* (pp. 606–611). Mahwah NJ: Erlbaum.
- Doumas, L. A. A. & Hummel, J. E. (2010). A computational account of the development of the representations underlying object recognition. *Cognitive Science*, 34, 698–712.
- Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, 115, 1–43.
- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2), 179–211.
- Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41, 1–63.
- Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture. *Cognition*, 28, 3–71.
- Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, 19, 141–205.
- Gentner, D., & Forbus, K. (2011). Computational models of analogy. *WIREs Cognitive Science*, 2, 266–276.
- Gentner, D., Ratterman, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology*, 25, 524–575.
- Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, 15, 1–38.
- Halford, G. S., Wilson, W. H., Guo, J., Gayler, R. W., Wiles, J., & Stewart, J. E. M. (1994). Connectionist implications for processing capacity limitations in analogies. In K. J. Holyoak & J. A. Barnden (Eds.), *Advances in connectionist and neural computation theory: Vol. 2: Analogical connections* (pp. 363–415). Norwood, NJ: Ablex.
- Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Brain and Behavioral Sciences*, 21, 803–864.
- Helie, S., & Sun, R. (2010). Incubation, insight, and creative problem solving: A unified theory and a connectionist model. *Psychological Review*, 117, 994–1024.
- Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press.
- Holyoak, K. J., & Hummel, J. E. (2000). The proper treatment of symbols in a connectionist architecture. In E. Deirich & A. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines* (pp. 229–263). Mahwah, NJ: Erlbaum.
- Holyoak, K. J., & Hummel, J. E. (2001). Toward an understanding of analogy within a biological symbol system. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 161–195). Cambridge, MA: MIT Press.
- Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought*. Cambridge, MA: MIT Press.
- Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences USA*, 79, 2554–2558.
- Hummel, J. E. (2000). Where view-based theories break down: The role of structure in human shape perception. In E. Deirich & A. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines* (pp. 157–185). Mahwah, NJ: Erlbaum.
- Hummel, J. E. (2001). Complementary solutions to the binding problem in vision: Implications for shape perception and object recognition. *Visual Cognition*, 8, 489–517.
- Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, 99, 480–517.
- Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104, 427–466.
- Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, 110, 220–263.
- Hummel, J. E., Holyoak, K. J., Green, C., Doumas, L. A. A., Devnich, D., Kittur, A., & Kalar, D. J. (2004). A solution to the binding problem for compositional connectionism. In S. D. Levy & R. Gayler (Eds.), *Compositional connectionism in cognitive science: Papers from the AAAI fall symposium* (pp. 31–34). Menlo Park, CA: AAAI Press.
- Joanisse, M., & Seidenberg, M. S. (1999). Impairments in verb morphology following brain injury: A connectionist model. *Proceedings of the National Academy of Sciences USA*, 96, 7592–7597.
- Kanerva, P. (1998). *Sparse distributed memory*. Cambridge, MA: MIT Press.
- Kellman, P. J., Burke, T. & Hummel, J. E. (1999). Modeling perceptual learning of abstract invariants. In *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society* (pp. 264–269). Mahwah, NJ: Erlbaum.
- Krawczyk, D. C., Holyoak, K. J., & Hummel, J. E. (2005). The one-to-one constraint in analogical mapping. *Cognitive Science*, 29, 29–38.
- Kubose, T. T., Holyoak, K. J., & Hummel, J. E. (2002). The role of textual coherence in incremental analogical mapping. *Journal of Memory and Language*, 47, 407–435.
- Leech, R., Mareschal, D., & Cooper, R. P. (2008). Analogy as relational priming: A developmental and computational perspective on the origins of a complex cognitive skill. *Behavioral and Brain Sciences*, 31, 357–378.
- Li, S. C., Lindenberger, U., & Sikstrom, S. (2001). Aging cognition: From neuromodulation to representation to cognition. *Trends in Cognitive Sciences*, 5, 479–486.
- Love, B. C. (1999). Utilizing time: Asynchronous binding. *Advances in Neural Information Processing Systems*, 11, 38–44.
- Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive psychology*, 37(3), 243–282.
- Marcus, G. F. (2001). *The algebraic mind: Integrating connectionism and cognitive science*. Cambridge, MA: MIT Press.
- Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.
- Marshall, J. A. (1995). Adaptive pattern recognition by self-organizing neural networks: Context, uncertainty, multiplicity, and scale. *Neural Networks*, 8, 335–362.
- McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, 114, 159–197.
- McClelland, J. L., Rumelhart, D. E., & The PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2). Cambridge, MA: MIT Press.

- McCulloch, W. A., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5, 115–133.
- Metcalfe, J. (1990). Composite Holographic Associative Recall Model (CHARM) and blended memories in eyewitness testimony. *Journal of Experimental Psychology: General*, 119, 145–160.
- Morrison, R. G., Krawczyk, D. C., Holyoak, K. J., Hummel, J. E., Chow, T. W., Miller, B. L., & Knowlton, B. J. (2004). A neurocomputational model of analogical reasoning and its breakdown in frontotemporal lobar degeneration. *Journal of Cognitive Neuroscience*, 16, 260–271.
- Munakata, Y., & O'Reilly, R. C. (2003). Developmental and computational neuroscience approaches to cognition: The case of generalization. *Cognitive Studies*, 10, 76–92.
- Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.
- O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation*, 8, 895–938.
- O'Reilly, R. C. (2006). Biologically based computational models of high-level cognition. *Science*, 314, 91–94.
- O'Reilly, R. C., & Busby, R. S. (2002). Generalizable relational binding from coarse-coded distributed representations. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems (NIPS)* 14 (pp. 75–82). Cambridge, MA: MIT Press.
- O'Reilly, R. C., Busby, R. S., & Soto, R. (2003). Three forms of binding and their neural substrates: Alternatives to temporal synchrony. In A. Cleeremans (Ed.), *The unity of consciousness: Binding, integration, and dissociation* (pp. 168–192). Oxford, England: Oxford University Press.
- Page, M. (2000). Connectionist modelling in psychology: A localist manifesto. *Behavioral and Brain Sciences*, 23, 443–512.
- Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008) Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, 31(2), 109–178.
- Pierce, C. S. (1879/1903). Logic as semiotic: The theory of signs. In J. Buchler (Ed.), *The philosophical writings of Pierce* (1955) (pp. 98–119). New York: Dover Books.
- Plate, T. (1991). Holographic reduced representations: Convolution algebra for compositional distributed representations. In J. Mylopoulos & R. Reiter (Eds.), *Proceedings of the 12th International Joint Conference on Artificial Intelligence* (pp. 30–35). San Mateo, CA: Morgan Kaufmann.
- Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.
- Rogers, T. T., & McClelland, J. L. (2008). Precis of semantic cognition, a parallel distributed processing approach. *Behavioral and Brain Sciences*, 31, 689–749.
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408.
- Ross, B. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 629–639.
- Rumelhart, D. E., McClelland, J. L., & The PDP Research Group. (1986). *Parallel distributed processing: Explorations in the micro-structure of cognition* (Vol. 1). Cambridge, MA: MIT Press.
- Sandhofer, C. M., & Doumas, L. A. A. (2008). Order and presentation effects in learning categories. *Journal of Cognition and Development*, 9, 194–221.
- Shastri, L. (2002). Episodic memory and cortico-hippocampal interactions. *Trends in Cognitive Science*, 6, 162–168.
- Shastri, L., & Ajjanagadde, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences*, 16, 417–494.
- Sheya, A., & Smith, L. B. (2006). Perceptual features and the development of conceptual knowledge. *Journal of Cognition and Development*, 7(4), 455–476.
- Smith, L. B. (1984). Young children's understanding of attributes and dimensions. *Child Development*, 55, 363–380.
- Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46, 159–216.
- Son, J. Y., Doumas, L. A. A., & Goldstone, R. L. (2010). When do words promote analogical transfer? *Journal of Problem Solving*, 3, 52–92.
- St. John, M. F. (1992). The Story Gestalt: A model of knowledge-intensive processes in text comprehension. *Cognitive Science*, 16, 271–302.
- St. John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, 46, 217–257.
- Sun, R. (2009). Motivational representations within a computational cognitive architecture. *Cognitive Computation*, 1, 91–103.
- Sun, R., Slusarz, P., & Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: A dual-process approach. *Psychological Review*, 112, 159–192.
- Taatgen, N. A., & Anderson, J. R. (2008). ACT-R. In R. Sun (Ed.), *Constraints in cognitive architectures* (pp. 170–185). Cambridge, UK: Cambridge University Press.
- Tesar, B., & Smolensky, P. (1994). Synchronous-firing variable binding is spatio-temporal tensor product representation. *Proceedings of the 16th Annual Conference of the Cognitive Science Society*. Atlanta, GA. August.
- Viskontas, I. V., Morrison, R. G., Holyoak, K. J., Hummel, J. E., & Knowlton, B. J. (2004). Relational integration, inhibition and analogical reasoning in older adults. *Psychology and Aging*, 19, 581–591.
- von der Malsburg, C. (1981). *The correlation theory of brain function* (Internal Report No. 81–82). Goettingen, Germany: Department of Neurobiology, Max-Planck-Institute for Biophysical Chemistry.
- von der Malsburg, C. (1999). The what and why of binding: The modeler's perspective. *Neuron*, 24, 95–104.
- Wharton, C. M., Holyoak, K. J., & Lange, T. E. (1996). Remote analogical reminding. *Memory and Cognition*, 24, 629–643.

Neurocognitive Methods in Higher Cognition

Robert G. Morrison and Barbara J. Knowlton

Abstract

The methods of cognitive neuroscience, notably functional neuroimaging and cognitive neuropsychology, are becoming increasingly important in efforts to understand the processes responsible for human higher cognition. Given the complexity of human thinking and reasoning, it is frequently the case that multiple theories can explain behavioral results. By utilizing the constraint of neural plausibility, some of these possibilities can be eliminated. These tools are thus beginning to help us to understand how thinking and reasoning actually occur in the brain. In this chapter we discuss a number of the techniques most frequently used to investigate higher cognition, including cognitive neuropsychology, scalp electroencephalography (EEG), functional magnetic resonance imaging (fMRI), and transcranial magnetic stimulation (TMS). We briefly survey a number of examples of how these techniques have contributed to our understanding of higher cognition, particularly the functions of the human prefrontal cortex.

Key Words: neuropsychology, neuroimaging, functional magnetic resonance imaging (fMRI), electrophysiology, event-related potentials (ERPs), prefrontal cortex

Introduction

For the past 40 years, neuroscience methods have played an increasingly important role in the study of cognition. It is now commonplace for cognitive scientists to connect cognitive processes to their underlying neural substrates. The explosive growth in the field of cognitive neuroscience, particularly in perception and memory, is blurring distinctions between cognitive psychology and neuroscience. Neuroscientists are now recognizing that higher cognition, including the study of thinking and reasoning, are also tractable areas for research, which could greatly benefit from attention to the constraint of neural plausibility.

The development of the field of cognitive neuroscience is a natural consequence of the fact that “cognition is what the brain does.” Recent years have seen unprecedented development in both the study of cognition and of the brain. The development

of neuroimaging techniques, chiefly functional magnetic resonance imaging (fMRI), has clearly accelerated this convergent growth. Because methodological developments have fueled advances in the cognitive neuroscience approach, this chapter is organized in terms of how different methodologies are informing the field.

Findings in cognitive neuroscience fall into two basic categories. In one category, researchers elucidate brain-behavior relationships; that is, they assign cognitive functions to specific brain regions or circuits. In the other category, neuroscience data are brought to bear in order to constrain cognitive theories, or they are used to provide a resolution between two theories that are both plausible based on behavioral data alone. This second category of findings is typically of more interest to cognitive scientists; however, the famous 19th-century neurologist Bernhard Von Gudden was wise to caution,

“Faced with an anatomical fact proven beyond doubt, any physiological result that stands in contradiction to it loses all its meaning... So, first anatomy and then physiology; but if first physiology, then not without anatomy” (as cited by Brodmann, 2006, p. 262). Thus, understanding the functional neuroanatomy of the brain is the first step in determining how the physical matter of the nervous system gives rise to human thought. Ultimately, this is one of the fundamental questions in life science. Thus, cognitive neuroscience is useful in that it provides additional methods for cognitive science, but it is also an important pursuit in its own right.

Methods of Cognitive Neuroscience

Building upon a long history of work in cognitive neuropsychology,¹ the methods of cognitive neuroscience are constantly evolving. In addition to functional neuroimaging techniques sensitive to the temporal dynamics or spatial localization of cognitive processes, researchers have also made extensive use of computational modeling to capture brain network architecture or functions, recently augmented by the methods of cognitive neurogenetics (see Green & Dunbar, Chapter 7). Here we introduce several of the techniques currently being used to study higher cognition, and we provide examples of their use (see Table 6.1 for a summary of methods).

Cognitive Neuropsychology

While modern “cognitive neuroscience” may have officially begun with the coining of the term in the late 1970s by Michael Gazzaniga and George Miller (D’Esposito, 2010), the precursors of this field can be traced to 19th-century studies of brain-damaged patients. The great controversy of the time was between *localizationism*, the view that specific cognitive functions could be ascribed to particular brain regions, versus an *aggregate field theory*, according to which cognitive abilities are distributed throughout the neocortex. Under the first view, restricted damage to specific brain regions should disrupt specific cognitive processes while leaving others intact. Under the second view, the extent of damage to the brain is more important than the location of damage, with all cognitive functions proportionately affected by damage. Some of the most compelling data from this period arguing for localizationism came from two patients studied by Paul Broca (Lee, 1981). These patients (Leborgne and Lelong) became unable to speak more than a few words. After the death of each patient, Broca

examined their brains and determined that for both patients the language difficulties were due to damage in the left inferior frontal lobe, a region now named Broca’s area (see Fig. 6.1). Interestingly, subsequent work has shown that patients with lesions limited to Broca’s area do not actually exhibit the kind of profound deficits in language production described in Broca’s original cases. Recent examination of Leborgne and Lelong’s brains with modern methods in fact demonstrate that the damage was much more extensive than originally described (Dronkers, Plaisant, Iba-Zizen, & Cabanis, 2007). Nevertheless, these case studies indicated that a complex cognitive function like language production could be selectively affected by brain damage that generally spared other functions.

Broca’s dissociation-based approach (see Fig. 6.2), which looks for commonalities in spared and impaired function with associated brain damage across subjects, has continued to be used in modern cognitive neuroscience. The power of the cognitive neuropsychological approach is that it can tell us whether a specific brain region is *necessary* for a particular cognitive function, and whether remaining regions are *sufficient* to support functions that are spared (i.e., single dissociation). In addition, even in situations where the location of damage is unclear, much can be learned about the organization of cognition by studying how it breaks down. For example, the apparent relative sparing of language comprehension in patients Leborgne and Lelong, despite their severe difficulties with language production, support models of language in which these abilities are independent. Another example is the distinction between declarative and procedural memory systems. The strongest evidence that memory is organized into such systems comes from the fact that amnesic patients are able to learn skills and procedures normally despite extremely poor memory for practice episodes (Cohen & Squire, 1980; Knowlton, Mangels, & Squire, 1996).

The neuropsychological approach depends on a careful analysis and characterization of behavior (see Feinberg & Farah, 2003). Dissociations between cognitive functions can be interpreted in different ways depending on the underlying psychological theory. Although a straightforward interpretation of the findings from Broca’s patients suggested a double dissociation between language production and comprehension, subsequent research has shown that similar patients have difficulty with comprehension based on grammar. It appears more accurate to describe these patients as “agrammatic.” To

Table 6.1. Summary of Experimental Methods in Cognitive Neuroscience

Temporal Resolution						
Method	Temporal Lag	Sampling Rate	Spatial Resolution	Advantages	Disadvantages	Methods and Examples
Cognitive neuropsychology (Behavioral Neurology)	NA	NA	10^{-3} – 10^{-2} m	Can access whether a particular brain region is necessary for a particular cognitive function; can be combined with structural and/or functional imaging and be used to test computational models	Depends on the “experiments of nature”; requires precise characterization of damage (structural imaging) and careful testing to establish dissociations	Feinberg & Farah (2003); Waltz et al. (1999); Morrison et al. (2004); Huey et al. (2009); Sonty et al. (2007)
Computed axial tomography (CT)	NA	Can be repeated over days to years	10^{-2} – 10^{-1} m	Excellent vascular detail; availability in all major medical/research centers	Poor detail except for vascular structures	
Structural magnetic resonance imaging (MRI, cortical thickness, VBM)	NA	Can be repeated over days to years	10^{-3} – 10^{-1} m	Excellent gray/white matter detail; can be used to measure cortical thickness or used with techniques like voxel-based morphometry (VBM) to make group comparisons; can be correlated with behavior	Claustrophobic; cannot currently be used in individuals with medical implants; difficult to use with infants and children	Raichle (1994); Fischl & Dale (2000); Ashburner & Friston (2000); Dumontheil et al. (2010); Rosen et al. (2002)
Diffusion tensor imaging (DTI)	NA	Can be repeated over days to years	10^{-3} – 10^{-1} m	Used to measure white matter integrity (DTI); can be correlated with behavior	Expensive; claustrophobic; difficult to use with infants and children	Filler (2009); Rogalski et al. (2009)
Single- or multi-unit recording	Instant	ms	10^{-5} – 10^{-3} m	Superior spatial and temporal resolution	Invasive! (animals and very selective brain-damaged humans); moderate startup and per-participant costs in animals	Humphrey & Schmidt (1990); Fuster & Alexander (1971); Cromer, Roy, & Miller (2010)

(Continued)

Table 6.1 Continued

Temporal Resolution						
Method	Temporal Lag	Sampling Rate	Spatial Resolution	Advantages	Disadvantages	Methods and Examples
Electrocorticography (ECOG)	Instant	ms	10^{-3} – 10^{-2} m	Good spatial and temporal resolution	Invasive! (very selective brain-damaged humans); high startup and per-participant costs	Miller et al. (2007); Canolty et al. (2006)
Scalp electroencephalography (EEG/ERP/ERO)	Instant	ms	10^{-2} – 10^{-1} m?	Noninvasive; excellent temporal resolution; low cost; can be used with children and infants	Undefined spatial resolution; sensitive to movement and eye artifacts	Luck (2005); Sauseng & Klimesch (2008); Knight et al. (1989); Vogel et al. (2005); Jung-Beeman et al. (2004); Kounios et al., (2006)
Positron emission tomography (PET)	2 m	?	10^{-2} – 10^{-1} m	Can be used either structurally or functionally; can be tuned to different metabolic processes; limited availability	Invasive (radiation); very expensive per-participant costs	Raichle (1983); Cabeza & Nyberg (2000); Villemagne et al. (2011)
Functional magnetic resonance imaging (fMRI)	1–2 s	250 ms - 1 s	10^{-3} – 10^{-1} m	Noninvasive; available at major medical/research centers; excellent spatial resolution	Expensive; loud and claustrophobic; cannot currently be used in individuals with medical implants; limited movement, poor participant contact, high per-participant cost	Brown, Perthen, Liu, & Buxton (2007); Raichle & Mintun (2006); Cabeza & Nyberg (2000); Kroger et al. (2002); Goel & Dolan (2004); Monti, Parsons, & Osherson (2009); Sonty et al. (2007)
Magnetoencephalography (MEG)	Instant	ms	10^{-2} – 10^{-1} m?	Noninvasive, provides high temporal resolution and moderate spatial resolution in one technique; can be used with children and infants	Not widely available in United States; high startup costs; moderate operating costs—much lower than fMRI	Hansen, Kringselbach, & Salmelin (2010); Ciesielski, Ahlfors, Bedrick, Kerwin, & Hamalainen (2010)

Near infrared spectroscopy (NIRS)	1–2 s	ms	10^{-3} – 10^{-2} m	Noninvasive; low-startup and low per-participant costs; can be used with infants and young children	Can only be used with cortical surface structures near scalp	Villringer & Chance (1997); Tsujii & Watanabe (2010)
Event-related optical signal (EROS)	100 ms	ms	10^{-3} – 10^{-2} m	Noninvasive; low-startup and low per-participant costs; can be used with infants and young children	Can only be used with cortical surface structures near scalp; very limited availability	Gratton et al. (1997)
Transcranial magnetic stimulation (TMS)	Instant	NA	10^{-3} – 10^{-2} m	Noninvasive virtual lesion method; relatively low cost but usually done with fMRI	Cannot be used with people with tendency to have seizures	Pascual-Leone, Bartres-Faz, & Keenan (1999); Tsujii et al. (2010)

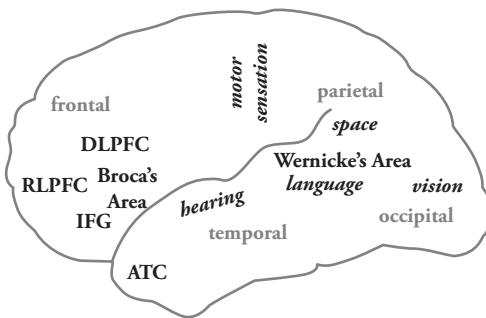


Fig. 6.1 Functional neuroanatomy in higher cognition. The human cerebral cortex is traditionally divided into four lobes based on neuroanatomical landmarks (e.g., the horizontal fissure). These lobes are then frequently further partitioned with respect to smaller anatomical divisions (e.g., prefrontal cortex [PFC], rostralateral PFC [RLPFC], dorsolateral PFC [DLPFC], inferior frontal gyrus [IFG], anterior temporal cortex [ATC]). However, sometimes brain regions are labeled with the name of the scientist who originally ascribed their function (e.g., Broca's or Wernicke's area), the type of information they process (e.g., auditory, sensation, visual), or by their dominant function (e.g., motor movement, language, spatial processing).

the extent that they can produce language, they will produce content words, but not function words that would indicate the use of grammar (Kean, 1977). Thus, more extensive examination of behavioral findings along with the development of new theoretical perspectives can lead to reinterpretations of neuropsychological data.

In the domain of thinking and reasoning, patients with frontal lobe damage are of interest because of the clear involvement of the frontal lobes in complex cognition (see also Holyoak, Chapter 13). Focal lesions to the frontal lobes (e.g., following a stroke) are very common; however, they are likely to be unilateral and restricted, making it likely that spared regions can take over lost functions. In addition,

the location and the extent of the lesions often vary between patients, making it difficult to generalize across cases (see Duncan & Owen, 2000). In investigations of neural mechanisms of reasoning, a great deal of attention has recently focused on patients with frontotemporal lobar degeneration (FTLD; e.g., Huey et al., 2009; Krawczyk et al., 2008; Morrison et al., 2004; Waltz et al., 1999; Zamboni, Huey, Krueger, Nichelli, & Grafman, 2008).² Although these patients are much rarer than those with focal frontal lobe damage, their damage is more encompassing of prefrontal cortex and can thus provide a good picture of its role in higher cognition.

FTLD can present with different clusters of symptoms depending on the regions of initial involvement (Mesulam, 2007; Miller, 2007). Although the disease ultimately progresses to involve multiple brain regions, in early stages it can affect specific regions of the brain more selectively. Those patients in the frontal-variant (also referred to as behavioral-variant) category exhibit executive problems early on in the disease course. Another group of patients, with early involvement in the left temporal lobe, exhibit deficits in semantic knowledge (i.e., semantic dementia). These two patient groups have been studied in the context of thinking and reasoning, as they exhibit a contrasting set of deficits. Using the technique of voxel-based morphometry to quantify the regional extent of damage, it is possible to correlate the extent of degeneration with specific cognitive abilities (Huey et al., 2009; Rosen et al., 2002). These studies have generally shown correlations between degeneration in the anterior frontal lobe and standardized tests of problem solving, whereas degeneration in other regions is strongly associated with other aspects of higher cognition. For example, degeneration in the dorsolateral prefrontal cortex

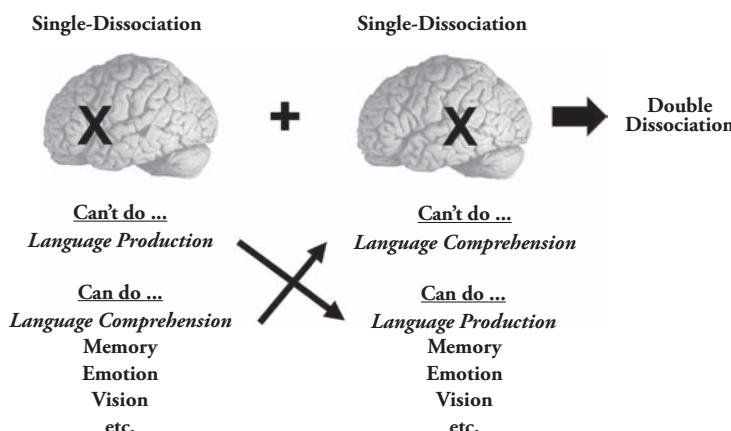


Fig. 6.2 Cognitive neuropsychology. Patient-based methods are typically based on the logic of single or double dissociations.

(DLPFC) is correlated with apathy, while degeneration in the right posterior temporal lobe is related to the patient's insight into his or her own behavioral problems (Zamboni et al., 2008).

Despite the widespread deficits in executive function shown by frontal-variant patients, there is also evidence that there is some selectivity in these deficits. For example, in tests of inductive and deductive reasoning, frontal-variant patients can perform within normal range when solving the problem requires only a single relation to be kept in mind (e.g., a Raven's Progressive Matrix problem that can be solved by considering a change across a single dimension, such as a shape getting larger across a row). However, when the subject must consider multiple relations simultaneously to find the answer (e.g., the shape is getting larger across rows and darker across columns), frontal-variant patients perform at chance levels. These results suggest that relational integration is one of the key contributions to reasoning made by the prefrontal cortex (Waltz et al., 1999).

In contrast to the performance of frontal-variant patients, those with temporal lobe involvement do not appear to have a deficit in relational integration. Rather, these patients exhibit deficits in using semantic knowledge to make inferences (Krawczyk et al., 2008; Morrison et al., 2004), while they are often able to reason about materials for which semantic content is not important, such as the Raven's Progressive Matrices test (Waltz et al., 1999). These results suggest that different neural circuits subserve reasoning about familiar versus unfamiliar domains. One interpretation of these findings is that reasoning based on familiar information necessarily uses the semantic knowledge system—in other words, it is difficult to reason with familiar entities as if they are arbitrary (see Evans, Chapter 8). In contrast, when there is an arbitrary relationship between items, a system for applying formal logical rules is engaged. Based on the results from FTLD patients as well as neuroimaging studies, interactions between frontal and temporal lobes are crucial for reasoning when semantic knowledge is relevant (Krawczyk et al., 2008, Luo et al., 2003; Morrison et al., 2004), whereas frontoparietal circuitry plays a major role in applying formal rules of logic to arbitrary relations (Hinton, Dymond, von Hecker, & Evans, 2010; Noveck, Goel, & Smith, 2004). Because the parietal lobes remain relatively intact in FTLD, these patients are able to solve deductive and inductive reasoning problems with

arbitrary content, particularly if they involve only a limited number of relations. With arbitrary relations, patients with frontal involvement perform poorly when they must consider multiple relations simultaneously in working memory (Waltz et al., 1999) or when activated semantic information competes during reasoning (Krawczyk et al., 2008; Morrison et al., 2004). The study of patients with FTLD has provided strong support for the idea that reasoning with semantically meaningful and arbitrary relations involves different neural circuitry.

Neuroimaging

Cognitive neuropsychology continues to yield new insights into the dissociable processes within the domain of reasoning. However, while this approach emphasizes dissociation, other approaches are needed to understand how regions act cooperatively. Neuroimaging approaches have a distinct advantage over cognitive neuropsychology in that they are noninvasive and involve relatively large numbers of healthy subjects, allowing greater generalizability of their findings. As cognitive neuropsychology depends on “experiments of nature,” there is unavoidable variability between patients in terms of lesion site or course of illness, which can make it difficult to make inferences about those regions responsible for observed deficits. Neuroimaging approaches, on the other hand, allow one to glimpse the intact brain at work. Neuroimaging methods vary with respect to whether they measure the structure or function of the nervous system; however, results from either type of method can be correlated with behavioral measures to provide valuable information about cognitive function. Functional methods can in turn either directly or indirectly measure neural activity. For instance, scalp electroencephalography (EEG) directly measures changes in voltage resulting from firing neurons, whereas functional magnetic resonance imaging (fMRI) indirectly measures neuronal activity by measuring increased blood flow to the area of the brain recently active. A further distinction is between methods that focus on spatial localization (e.g., fMRI) versus those that provide information on temporal dynamics (e.g., EEG).

STRUCTURAL NEUROIMAGING

Although neurologists are able to look at X-rays of the head to see damage to the skull, X-rays are inadequate to image soft tissue such as the brain. In 1974, brain imaging took a huge step forward with the development of computer axial tomography (CT or

CAT scan). This enhanced three-dimensional X-ray was able to resolve gray and white matter as well as blood and cerebrospinal fluid. As a consequence, neurologists were able to see the damage caused by tumors and different types of irregular blood flow (e.g., ischemia and aneurysms). This greatly facilitated cognitive neuropsychology because researchers did not have to wait until a patient died to know what areas of their brain were damaged. The development of magnetic resonance imaging (MRI), with its greater white/gray contrast and finer spatial resolution, allowed for precise volumetric measurement of different brain structures (Raichle, 1994). The state of the art in structural MRI allows researchers to measure cortical thickness (Fischl & Dale, 2000) or white matter integrity (Filler, 2009), correlating it with various types of behavior and even developmental change.

Comparisons can also be made across groups using techniques such as voxel-based morphometry (VBM; Ashburner & Friston, 2000), which allows researchers to spatially normalize brain images into a common stereotactic space, and then make voxel-by-voxel comparisons of the local concentration of gray matter between groups. This technique is particularly useful for demonstrating the similarities of cortical damage in different patients groups. For instance Rosen et al. (2002) used VBM to characterize structural differences in various subtypes of FTLD. Variability in a particular brain region can also be correlated with behavioral changes (see Huey et al., 2009).

In a recent developmental study of reasoning, Dumontheil, Houlton, Christoff, and Blakemore (2010) tested a large group of children using a relational reasoning task that shows major behavioral changes during adolescence. Using structural MRI with both cortical thickness and VBM analyses, they found significant reductions in gray matter but not white matter volume during adolescence in areas of prefrontal cortex functionally involved in relational reasoning. These results suggested that improvements in relational reasoning can be the result of decreases in the number of synapses, allowing for an increase in effective connectivity between brain regions necessary for reasoning.

Although structural MRI allows for excellent contrast between gray and white material, it does not directly measure the integrity of the tissue. In contrast, diffusion tensor imaging (DTI) can be used to appraise the integrity of white matter by using different settings during MRI image capture

(Filler, 2009). The analysis procedure appraises the characteristics of water in the tissue. Clinically, these methods have been used to diagnosis multiple sclerosis and recently have also been potentially helpful in detecting early stages of Alzheimer's disease (Rogalski et al., 2009). Importantly, white matter (the axons of myelinated neurons) connects different regions of the brain, and it is critical for both working memory (frontal/parietal network) and language (frontal/temporal network). Thus, it is likely that differences in connectivity as measured by DTI may be useful for appreciating individual differences in thinking, as well as development.

ELECTROPHYSIOLOGICAL FUNCTIONAL NEUROIMAGING

Single- and Multi-Unit Recording

Our understanding of many basic cognitive functions has been profoundly aided by studies using electrophysiological methods with nonhuman animals. In these studies microelectrodes are inserted into precise locations in the brain and can be used to directly record the firing of either single (i.e., single-unit recording) or small groups (i.e., multi-unit recording) of neurons (see Humphrey & Schmidt, 1990). This technique produces results with excellent temporal and spatial resolution—we know exactly when and where neurons are firing when a particular cognitive process is engaged. This information is particularly useful when cognitive processes can be clearly defined. For instance, if we would like to know what neurons respond to the spatial frequency of a visual pattern, we can place electrodes in various regions in primary visual cortex and locate neurons that fire to a particular spatial frequency, but not other frequencies. Similarly, we can ask whether there are cells that are specifically sensitive to face stimuli and not other complex visual objects.

Unfortunately, it is frequently difficult to isolate cognitive processes underlying thinking and reasoning with this same degree of precision. As Penn and Povinelli argue in Chapter 27, it is very likely that humans differ from even their nearest primate relatives in the nature of relational representations and processing. Despite these limitations, electrophysiology in the macaque monkey has provided some insight into the functions of the prefrontal cortex for higher cognition. Much of higher cognition, particularly System II or explicit processing (see Evans, Chapter 8), depends heavily on the working memory system for the maintenance

and manipulation of information (see Morrison, 2005). Seminal studies in the macaque by Fuster and Alexander (1971) demonstrated that neurons in prefrontal cortex selectively fire during the delay in delayed match-to-sample tasks, in which a monkey is required to match a target to a previously displayed sample object shown before a brief delay. In an elegant study using cortical cooling to temporarily deactivate connective fiber tracts, Fuster, Bauer, and Jervey (1985) went on to demonstrate that prefrontal neurons were not simple buffers for the temporary storage of information, but rather were responsible for maintaining the activity of neurons in posterior cortex, which actually coded for the information being maintained. Consistent with Cowan's (1995) conception of working memory as a process of selective attention, Fuster's finding illustrates one central "truth" of higher cognition—brain regions dynamically collaborate to accomplish complex processes.

Electrocorticography

The invasive nature of single- and multi-unit recording typically prevents its use in humans. One exception is in patients with intractable epilepsy. These patients frequently have surgery to remove parts of the brain (usually in the temporal lobe) responsible for initiating or propagating seizures. In preparation for surgical resection, patients frequently have electrodes placed directly on the brain under the skull and dura matter. In some cases electrodes capable of single- or multi-unit recording are even inserted into the cortex (i.e., depth electrodes). Patients are monitored for up to several weeks waiting for seizures to occur. During this period patients typically participate in a variety of cognitive studies, which allow researchers to correlate brain activity with function (see Miller et al., 2007). Although this provides an excellent opportunity for directly recording the activity of neurons in humans, caution must be exercised in interpreting results because of the pathology associated with epilepsy in these patients.

One recent example of how electrocorticography (ECOG) can be used to constrain cognitive models, demonstrated that different areas of cortex communicate with each other during behavioral tasks by precise timing of different populations of neurons firing at different frequencies (Canolty et al., 2006). It was hypothesized that these communication patterns may be regulated by GABAergic inhibitory neurons in the basal forebrain. In general, it

appears that local, domain-specific neural circuits communicate using high frequencies (e.g., *gamma*) while more distant cross-domain circuits use lower frequencies (e.g., *theta*; Canolty & Knight, 2010). These observations are consistent with those symbolic-connectionist accounts of higher cognition (see Doumas & Hummel, Chapter 5) that use temporal synchrony as a binding mechanism for representing explicit relational structures necessary for relational thinking and reasoning.

Scalp Electroencephalography

The ability to directly record electrical activity from the brain began long before the advent of single- and multi-unit recording. In 1929 Hans Berger first used electrodes on the scalp to record the summation of voltage changes associated with the firing of millions of neurons. This early method of neuroimaging showed rhythmic patterns associated with different states of consciousness. For instance, cycles of approximately 10 times per second (10 Hz, *alpha*) were detected in electrodes over the occipital lobe during sleep or periods when an individual's eyes were closed. The method of scalp electroencephalography (EEG) was used for many years to diagnose a number of different medical conditions, including epilepsy.

Event-Related Potentials

The true power of EEG for neuroimaging was not realized until computing advances allowed for large numbers of measurements to be summed (Galambos & Sheatz, 1962). The resulting event-related potentials (ERPs) can be time locked to particular events, such as the presentation of a stimulus, or a button press in response. Averaging many time-locked trials increases the signal-to-noise ratio, resulting in a smooth waveform with characteristic positive and negative peaks (see Fig. 6.3). Over thousands of experiments, researchers have associated many of these peaks with particular cognitive processes (For a detailed introduction to EEG/ERP methods, see Luck, 2005; Luck & Kappenman, 2012).

Typically, researchers compare ERPs from different within-subject conditions to isolate particular neural components; however, ERPs can also be used to compare different groups as well. Here we briefly consider two different studies showing the power of ERPs to elucidate the function of prefrontal cortex. In general, one important function of prefrontal cortex is to filter information. The filtering can either take the form of tonic gating, serving

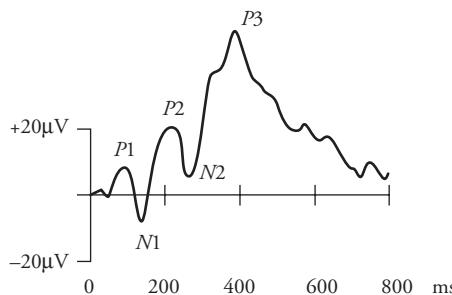


Fig. 6.3 Sample ERP. Averaging many trials from the same condition in an experiment improves the signal-to-noise ratio of time-locked EEG signal, resulting in an event-related potential with positive and negative peaks, which can sometimes be associated with particular cognitive processes.

to down-regulate sensory input from the outside world, and thereby avoiding distraction (Knight, Scabini, & Woods, 1989); or dynamic filtering to manage the contents of working memory for a current goal (Vogel, McCollough, & Machizawa, 2005). These are both essential aspects of executive functioning necessary for higher cognition that rely on prefrontal cortex.

Knight, Scabini, and Woods (1989; see also Knight & Graboweczyk, 1995) provided one early example of how ERPs can be used in conjunction with cognitive neuropsychology. They exposed patients with damage to DLPFC, or age- and education-matched control participants, to a sound recording of simple auditory clicks. These sounds produce positive deflections in the ERP within 30 ms of the click (P30). A group analysis revealed that patients and controls showed similar P30 ERPs when recorded from posterior electrode positions, whereas patients showed much more positive P30s than controls when the ERP was measured from parietal-, temporal-, or frontal-positioned electrodes. This pattern suggests that a functioning DLPFC serves to protect the higher cortical areas from sensory distraction by down-regulating signals.

A second ERP study investigating this issue explored how the ability to dynamically filter information can determine effective working-memory capacity. Numerous studies have linked individual differences in working-memory capacity with the ability to reason (Conway, Kane, & Engle, 2003). Vogel et al. (2005) used a delayed match-to-sample paradigm, similar to that used by Fuster in monkeys, in humans who were divided into two groups based on their individual working-memory spans (Conway et al., 2005). In Vogel et al.'s task participants had to remember the exact location and orientation

of several color bars over a delay. There were three different task conditions: (1) remember two bars of one color, (2) remember four bars of two different colors, or (3) remember two bars of one color and ignore two bars of a different color. Thus, the two latter conditions (2 and 3) had identical stimuli, with only the task differing. Vogel et al. isolated a contingent negative variation (CNV) in the EEG signal that began when the sample stimuli disappeared and persisted during the delay before participants were to respond to a target. The CNV thus appears to be an analog to the neural activity observed by Fuster using single-unit recording in nonhuman primates. Vogel and colleagues demonstrated that the CNV was modulated by working-memory load, with greater loads (remember-four) resulting in a more negative CNV than smaller loads (remember-two). However, they found that for people with low working-memory span, the remember-two and ignore-two condition produced the same CNV as the remember-four condition, while for high working-memory span people, the CNV for the remember-two and ignore-two condition looked like that for the remember-two condition. Thus, it appears that people with greater working-memory spans do not have greater working-memory capacities; rather, they simply manage the capacity they have more efficiently via dynamic filtering as regulated by prefrontal cortex.

Event-Related Oscillations

While the vast majority of EEG studies in cognitive neuroscience to date have used ERP analysis techniques, ERPs collapse the voltage data in such a way as to obscure the true nature of neuronal firing patterns. Clusters of neurons found in neural circuits tend to fire in oscillatory waves with characteristic frequencies. To capture this information, an increasing number of EEG studies in cognitive neuroscience have begun to employ what we term here event-related oscillation analyses (ERO; see Sauseng & Klimesch, 2008). These analyses are based on mathematical techniques (Fourier transform or wavelet analysis) that transform the recorded voltage changes into a frequency spectrum. Thus, researchers can estimate the relative populations of neurons firing at different rates. This is the type of analysis used in the ECOG study described previously (Canolty et al., 2006), but it can also be used with standard EEG recording as well.

Van Steenburgh et al. (Chapter 23) describe several different studies using time-frequency analysis to investigate the role of insight in problem

solving. In one study, Jung-Beeman et al. (2004) found two notable differences between problems that participants solved with or without insight. First, in problems solved with insight, they measured a sustained burst of low-alpha (around 10 Hz) EEG activity over the right parietal-occipital cortex beginning 1.5 seconds before the participant reported an answer. Low-alpha activity over visual cortex is understood to reflect visual sensory gating; thus, the brief deactivation of visual cortex may reflect a reduction of distracting sensory inputs needed in preparation for insight. Second, just 300 ms before answering, a burst of higher frequency gamma (30–80 Hz) activity over the right anterior temporal cortex was measured. The anterior temporal cortex is believed to be important for semantic integration, so it is likely that this burst is the signature for this cognitive process.

Although time-frequency analyses are frequently performed time-locked to stimuli or responses, they can also be used as a more general appraisal of cognitive state. In addition to the time-locked findings mentioned in the previous paragraph, Kounios et al. (2006) found that low-alpha EEG activity (8–10 Hz) prior to problem solving predicted whether a participant would solve the ensuing problem with insight, with greater alpha signifying a greater likelihood of a solution with insight. Likewise, many researchers have found that frontal asymmetries in low-alpha EEG activity also predict trait tendencies with respect to a general withdrawal or avoidance system (Davidson, 1993).

This type of analysis will likely prove increasingly useful as researchers in higher cognition seek to model the temporal dynamics of neuronal circuits. Changes in these dynamics have already been shown to provide parsimonious accounts for changes in reasoning associated with cognitive development (Morrison, Doumas, & Richland, 2011), aging (Viskontas, Morrison, Holyoak, Hummel, & Knowlton, 2004), and brain damage (Morrison et al., 2004).

SPATIAL FUNCTIONAL NEUROIMAGING

Spatial functional neuroimaging techniques make it possible to locate regions of activity in the brain with a greater level of precision than is possible with noninvasive electrophysiological techniques, such as EEG (see Table 6.1). Despite advances in source localization, there is nevertheless some ambiguity as to the brain regions that are contributing to an EEG signal because voltage is measured on the scalp, far from the many potential source generators. As a

result, a given pattern of electrical activity as measured on the scalp can arise from activation in many different areas of cortex. Techniques such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) can be used to link cognitive functions with specific locations in the brain (see Cabeza & Nyberg, 2000). Both of these techniques measure biophysical changes associated with metabolic activity and exploit the fact that neural activity requires energy. Thus, locations where there is more neural activity should be the source of greater metabolic activity. Thus, unlike EEG these techniques provide an indirect measure of neural activity; however, the greater spatial resolution they provide has been a key factor in the growth of cognitive neuroscience.

Positron Emission Tomography

In the 1980s, positron emission tomography (PET) became the first widely used functional neuroimaging technique providing adequate spatial resolution (see Raichle, 1983). In the most commonly used PET procedure, blood flow is measured using a radioisotope of water. Participants must be injected with this radioactive material, but the short half-life (about 123 seconds) renders it relatively safe. Blood flow (and thus the water isotope) is increased to regions that are metabolically active, and it can be detected by sensors outside the head. Complex mathematics implemented in a computer are used to build a three-dimensional map of where the changes in metabolic activity are located in the brain. A similar method, single-photon emission computed tomography (SPECT scan) is frequently used diagnostically by neurologists to characterize potential brain damage in patients showing abnormal behavior. Early studies using the PET procedure studied memory and perception. For example, regions that were determined to be involved in vision based on recordings from neurons in animal models were shown to have increased blood flow during visual tasks. Confirmatory studies such as these provided support for the validity of the PET technique.

Although fMRI has since eclipsed PET for the purpose of localizing cognitive functions to particular brain regions, PET remains an important technique for examining the role of neurotransmitters in cognitive function. These can be radiolabeled so that the uptake of the specific neurotransmitter in different brain regions can be localized (e.g., Okubo et al., 1997). Using specialized isotopes, PET has also found important diagnostic applications, including

the early detection of Alzheimer's disease (e.g., Villemagne et al., 2011).

Functional Magnetic Resonance Imaging

There are several advantages to the use of functional magnetic resonance imaging (fMRI) compared to PET. Rather than using radiation, fMRI measures the oxygenation level of blood in particular regions of the brain, and it exploits the fact that a magnetic field is disturbed differently based on the amount of oxygen in the blood. Because there is more oxygenated blood in regions where there is more neural activity, the BOLD (blood oxygenation level dependent) signal can be used to indirectly assess relative activity of brain regions (see Fig. 6.4). With PET, the exposure to radiolabeled compounds limits the frequency with which subjects can be tested, thus making it difficult to test changes across time or practice in individuals.

A major advantage of fMRI is that it has far greater spatial and temporal resolution than PET (Brown, Perthen, Liu, & Buxton, 2007; Raichle & Mintun, 2006). The BOLD signal can resolve changes in regions as small as a cubic millimeter using some specialized techniques. This enables researchers to test more precise anatomical hypotheses than are possible with PET or EEG measures. fMRI also is more temporally precise than PET, although the hemodynamic response still takes many seconds to develop. With fMRI, it is possible

to measure BOLD signal associated with individual trials (Glover, 1999). This advance enabled a much wider range of experimental paradigms than were possible with PET. For example, one could intermix trials with varying demands or eliminate trials in which the subject made an error. However, it remains the case that fMRI cannot provide the kind of millisecond temporal resolution that is possible with direct neural activity measures such as EEG or ECOG.

Ongoing work using fMRI methods may be able to gain leverage on issues that have proved difficult to resolve through behavioral experimentation alone. For example, there have been two general approaches to understanding how humans solve reasoning problems (see Evans, Chapter 8). According to one view, humans reason based on innate logical rules. According to another view (see Johnson-Laird, Chapter 9), we form a mental model of the premises of a problem and then identify solutions by scrutinizing the model. If we assume that mental logic rules are linguistic and propositional in nature (Braine, 1998; Rips, 1994), one might expect that reasoning should engage left hemisphere regions that are active during syntactic processing. By the mental models view, solving problems that involve relations that can be represented spatially (taller-than, better-than) should activate visuospatial regions in the right hemisphere (Knauff, Fangmeier, Ruff, & Johnson-Laird, 2003; Knauff & Johnson-Laird,

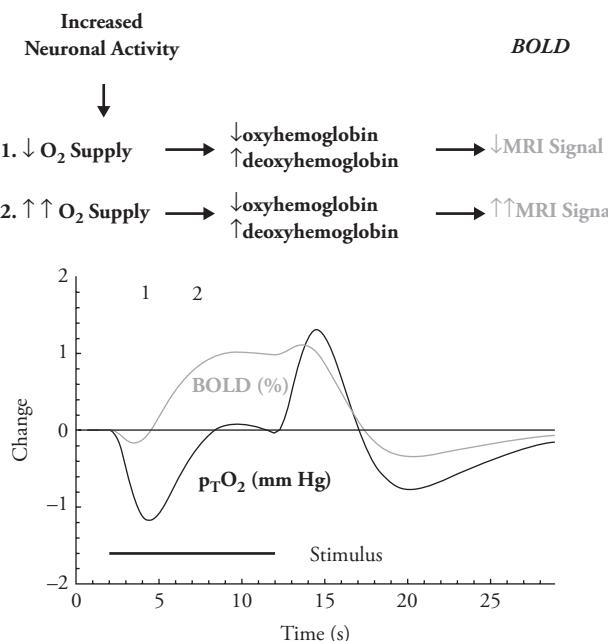


Fig. 6.4 Function magnetic resonance imaging (fMRI) is based on the hemodynamic response resulting from increases in metabolism in active neurons.

2002). At this point, existing fMRI data do not unequivocally support one view over the other, as the regions that are active during reasoning vary considerably depending on the task and the content of the reasoning problem. For example, solving syllogisms that are devoid of semantics such as (if P then Q, there is P, is there Q?) results in activation in left frontal-parietal pathways, whereas similar problems that use meaningful terms (If it is sunny, I ride my bike; I did not ride my bike today, is it sunny?) activate frontal and temporal lobe regions involved in semantic processing. Interestingly, when semantic knowledge or perceptual features of the problem are incongruous with the logical conclusion (If it is snowing, I ride my bike; I did not ride my bike today, is it snowing?), activation is also seen in right lateral midfrontal regions and the anterior cingulate (Goel & Dolan, 2003; Prado & Noveck, 2007), two areas that are implicated in cognitive control and conflict resolution. The fact that people often make errors when the logical conclusion of a problem is inconsistent with prior knowledge (the belief-bias effect, discussed later) can be viewed as a failure of engagement of these control circuits.

Overall, data from fMRI studies suggest that reasoning does not engage a dedicated neural circuit. Reasoning does not necessarily activate the right hemisphere regions involved in visuospatial processing, nor does it necessarily involve the same left hemisphere regions that are active during language processing (Kroger, Nystrom, Cohen, & Johnson-Laird, 2008; Noveck et al., 2004). However, these regions are engaged in some studies, supporting the possibility that rule application and visuospatial mental models can both support reasoning performance.

A view that is consistent with the mixed fMRI findings is that deductive reasoning involves a set of fractionated neural systems, each supporting different forms of relational reasoning (Goel, 2007). For example, although conditional reasoning (such as if-then problems) about unfamiliar information appears to rely on left frontoparietal regions, relational reasoning (e.g., Abe is to the left of Bill, and Bill is to the left of Charles; is Abe to the left of Charles?) relies in part on bilateral regions of the brain that have been implicated in visuospatial function, including the temporal-parietal-occipital junction (Goel & Dolan, 2004; Prado, Van Der Henst, & Noveck, 2010). Reasoning problems that involve semantically meaningful materials engage left temporal lobe storage sites for semantic information

(Goel, Buchel, Frith, & Dolan, 2000), consistent with the specific deficits in semantically meaningful problems exhibited by temporal-variant FTLD patients described earlier (Krawczyk et al., 2008; Morrison et al., 2004). The picture that appears to be emerging from these studies is that humans did not evolve a single reasoning system, but rather multiple systems suited to different problem domains are involved. Humans may have several reasoning systems at our disposal, including both rule-based mental logic and visuospatial mental models, and rely on whichever corresponding neural system is best suited for the problem (Prado et al., 2010).

Although fMRI studies to date do not seem to support any general theory of human reasoning, these studies together have consistently shown that deductive reasoning engages regions distinct from those involved in linguistic processing (Monti, Osherson, Martinez, & Parsons, 2007). While left hemisphere regions, including Broca's area in the inferior frontal gyrus (IFG), are often activated during deductive reasoning (Goel & Dolan, 2004), it is unclear if this is because language processing is required to understand the premises of a problem, or because these regions are generally important for syntactic processing that is common to both linguistic processing and logical reasoning. Based on fMRI approaches that allow the hemodynamic activity to be identified as the subject is processing different phases of the problem, it appears that although language areas in the frontal lobes are active for the initial interpretation of the premises, this activation quickly recedes to baseline, and nonlinguistic frontoparietal areas become active while the subject is actually solving the problem (Monti, Parsons, & Osherson, 2009). Thus, linguistic and logical rules used during reasoning appear to depend on distinct regions that occupy neighboring regions in left hemisphere, and fMRI data are at least making a good case that reasoning and thought can operate independently at the neural level. In addition, the existence of right hemisphere activation in several reasoning studies also suggests nonlinguistic processes may be engaged. For example, the involvement of right superior parietal lobule in these studies (Eslinger, et al. 2009; Goel & Dolan, 2003) is consistent with the idea that reasoning often involves the mental representation and manipulation of spatial information.

Although most fMRI studies of higher cognition have aimed to elucidate the various brain areas important for processing, there has been increasing

interest in understanding how these regions work together. In fact, the temporal dynamics of brain networks may be equally as important in higher cognition as simple activation. A dramatic example involving human language processing is a study by Sonty, Mesulam, Weintraub, Parrish, and Gitelman (2007), who used fMRI in conjunction with the cognitive neuropsychology approach. Patients with primary progressive aphasia (PPA; Mesulam, 2007; see also note 2), a neurodegenerative disease, show progressive loss of language functions, including the ability to name words and to appreciate the semantic relationships between concepts. Sonty et al. had patients with PPA, and age- and education-matched control participants, perform a semantic matching task (identify synonyms) and a letter-matching task (match nonword letter strings). Patients were less accurate than controls at the semantic task and slower on both tasks. Using fMRI, Sonty et al. found that several brain regions in the left hemisphere frequently associated with language (posterior fusiform gyrus, posterior superior temporal sulcus [Wernicke's area], IFG [anterior Broca's area], inferior parietal lobule, intraparietal sulcus, and ventral premotor cortex) were more active in the semantic than the letter task. Interestingly, these areas were not less active in patients than in controls, despite the poorer performance in patients. However, when Sonty et al. examined effective connectivity between these regions using dynamic causal modeling (DCM; Friston, Harrison, & Penny, 2003), a different story emerged. DCM uses a Bayesian decision algorithm to make estimates of how different brain regions affect each other via mono or polysynaptic interregional connections. Sonty et al. found that when patients were compared to controls, a significant decrease was found in connectivity between posterior superior temporal sulcus (Wernicke's area) and IFG (anterior Broca's area), and this decrease correlated with reduction in performance. Thus, in PPA it was not a change in activation *per se* that was responsible for changes in behavior, but rather the effectiveness of communication between areas. It is likely that connectivity patterns affect capabilities throughout higher cognition and may in part explain the dramatic individual differences and developmental patterns seen in cognitive abilities.

NEW METHODS FOR FUNCTIONAL NEUROIMAGING

Although EEG and fMRI currently dominate the neuroimaging methods used to study thinking

and reasoning, two new methods offer exciting possibilities for the future given their lower cost and better temporal resolution (relative to fMRI) and their superior spatial resolution (relative to EEG).

Magnetoencephalography

Like EEG, magnetoencephalography (MEG) directly measures neural activity; however, instead of measuring voltage, it utilizes the electromagnetic properties of the electrical charge associated with neurons to measure minute magnetic forces resulting from increases in neuronal firing (see Hansen, Krriegelbach, & Salmelin, 2010). According to Faraday's Law, magnetic forces exist perpendicular to current flow and as a result the magnetic forces do not spread as much as the voltage differences measured by EEG. Thus, the potential source generators for MEG signal are better constrained than those potentially responsible for EEG signal. Like EEG, MEG can be sampled with millisecond resolution. Also like EEG, the participant experience is quite noninvasive, allowing some participant movement during testing and good experimenter access to the participant. Like fMRI, MEG can also be used to examine when different brain regions tend to sequentially activate via DCM (Kiebel, Garrido, Moran, Chen, & Friston, 2009). As a result, MEG is rapidly becoming the preferred functional neuroimaging technique for imaging young children and infants. Thus, MEG will likely be important for understanding early changes in language and reasoning. For example, MEG has recently been used to compare 10-year-old children and adults in terms of the engagement of frontoparietal networks during working-memory processes (Ciesielski, Ahlfors, Bedrick, Kerwin, & Hamalainen, 2010). Even though the children and adults exhibited the same level of performance on an n-back task, differences in MEG signal in the two groups indicated that the children were relying on different neural mechanisms compared with adults.

Functional Optical Imaging

Functional near infrared spectroscopy (fNIRS) involves placing infrared light emitters and detectors on the scalp (see Villringer & Chance, 1997). The light penetrating the scalp is absorbed by oxy- and deoxyhemoglobin in the blood. The reflected wavelengths can then be measured, and the concentration changes as a result of neuronal activity can be estimated. Unlike fMRI, the NIRS response can measure both oxy- and deoxyhemoglobin,

with millisecond accuracy; however, just as with fMRI there is a delay in the beginning of the signal, because both techniques rely on measuring the hemodynamic response. NIRS equipment is relatively inexpensive, and unlike fMRI it costs very little to run the equipment. Thus, it is possible to run a larger number of participants in a study and thus potentially capture smaller behavioral effects. However, one problem with NIRS is that it can only detect neuronal activity in gray matter within approximately 1 cm of the scalp surface. While this limitation may prevent using NIRS to study the functions of subcortical structures, regions of the prefrontal cortex involved in working memory and reasoning are readily accessible using this technique. Like MEG, NIRS is being used in populations who cannot be tested in fMRI, including infants and young children.

An even more recently developed technique, event-related optical signal (EROS; Gratton et al., 1997), uses infrared light like fNIRS, but instead of measuring the hemodynamic response like fMRI it takes advantage of the light scattering qualities of neurons undergoing action potentials. Thus, EROS has a temporal latency (100 ms) closer to EEG and MEG than the multisecond latency of fMRI and fNIRS. Like NIRS, EROS can be sampled with millisecond frequency.

Tsujii and Watanabe (2010) recently used fNIRS to investigate the belief-bias effect, which has been explained using dual-process theory (Evans, Chapter 8). In this paradigm participants are asked to solve syllogisms in which the logical conclusions are either true (congruent) or not true (incongruent) about the world. The general finding is that participants are more likely to make reasoning mistakes when the logical conclusions are not factually correct—an effect that is enhanced when reasoning is speeded. Explanations based on dual-process theory argue that when resources (e.g., time) are limited, participants defer to a fast System I heuristic based on whether they believe the conclusion, rather than assessing whether it is logically correct. In contrast, when time is plentiful, participants use slower System II analytic processing. Tsujii and Watanabe used fNIRS to look at activity in right and left IFG, an area of the prefrontal cortex frequently associated with cognitive inhibitory functions, and an area they believed would be necessary to inhibit System I heuristic processes in order to choose the logically correct solutions when solving incongruent syllogisms. Their results were consistent

with this hypothesis. Specifically, they found that right IFG increased in activity on long-incongruent trials relative to short-incongruent trials, and that right IFG activity was correlated with the accuracy of individual participants on incongruent trials. These findings are broadly consistent with previous findings of right lateral prefrontal activation during incongruent trials (Goel & Dolan, 2003), and they extend these results in that they demonstrate a relationship between IFG activity and performance, and the ability to overcome belief-bias with time on each trial. The effects sizes in the Tsujii and Watanabe (2010) experiment were very modest, and they tested 48 participants in the study, something that would not have been feasible in an fMRI experiment due to expense.

Virtual “Lesions” Using Transcranial Magnetic Stimulation

Approaches such as EEG and fMRI seek to measure the output of the brain, and then correlate this activity with concurrent cognitive functions. Thus unlike cognitive neuropsychology, these approaches are unable to demonstrate whether this activity is necessary for the observed cognitive function. It is always possible that the activity that is measured is in fact supporting some incidental cognitive process. For example, activity associated with performance of a reasoning task could be related to incidental learning of the responses, or mental imagery, rather than the reasoning processes in question. Transcranial magnetic stimulation (TMS) allows the experimenter to alter function in a brain region and then measure the extent to which various cognitive functions are affected (see Pascual-Leone, Bartres-Faz, & Keenan, 1999). TMS relies on the fact that a magnetic field will induce an electric current orthogonal to the direction of the field. This induction is not impeded by the scalp or skull, so an electromagnetic coil held at the surface of the head can induce current in the brain tissue below. When TMS is administered as repetitive pulses, the effects of the current can be long lasting. In most experimental studies, TMS is used to disrupt function. However, if different stimulus intensities are used, it can also be used to enhance function. TMS given for extended periods over several sessions has been shown to be an effective treatment for some neurological and psychiatric disorders, as it appears to induce functional changes in activity (Wassermann & Lisanby, 2001).

Similar to the limitations of NIRS and EROS, a major disadvantage of the TMS technique is that

it is currently only possible to apply stimulation to regions on the surface of the brain that are accessible to the field generated by a coil. While thinking and reasoning certainly make use of subcortical structures, the frontoparietal network is clearly of great importance and accessible to TMS. For example, Tsujii et al. (2010) were able to follow up their NIRS study described previously using repetitive TMS to disrupt the region in the right IFG they found to be active when subjects needed to inhibit semantic-based heuristic processing to solve reasoning problems. As predicted, repetitive TMS applied to the right IFG interfered with performance on incongruent trials, thus enhancing the belief-bias effect.

TMS can be used in conjunction with neuroimaging techniques to examine whether regions of activation detected are in fact necessary for task performance. However, several important questions may be impossible to address with TMS, such as the interactions between cortical and subcortical structures in thinking (e.g., cortico-striatal loops) or the influence of emotion on thinking mediated by the amygdala.

Computational Modeling of Neural Systems

Computational modeling has greatly contributed to the development and testing of theories of thinking and reasoning at the computational and algorithmic levels of analysis (see, e.g., Doumas & Hummel, Chapter 5; Rips et al., Chapter 11; Buehner & Cheng, Chapter 12; Holyoak, Chapter 13; Koedinger and Roll, Chapter 40). Relatively little effort, however, has focused on understanding thinking and reasoning at the implementation level. Most implementation-level modeling of brain circuits has been based on connectionist architectures (e.g., Braver, Barch, & Cohen, 1999; O'Reilly, 2006), sometimes augmented with Bayesian decision rules. While these architectures and algorithms can capture many System I types of learning, they fail to capture many System II forms of thinking and reasoning that require explicit representation of relations (see Doumas & Hummel, Chapter 5). Notable exceptions include Anderson and colleagues (Anderson, Albert, & Fincham, 2005; Anderson et al., 2004) ACT-R model of problem solving, and Hummel and Holyoak's (1997, 2003) LISA model of relational reasoning.

ACT-R is a symbolic production system computer model of human cognition, including problem solving. ACT-R serves as a high-level programming language, including many basic assumptions about how cognition works but also allowing users to add

additional task-specific assumptions. ACT-R can make predictions for such indicators of task performance as response time or accuracy. Although ACT-R was not designed with brain architecture in mind, researchers have recently attempted to map several of ACT-R's basic functions to brain areas, and then to use this information to match fMRI activation patterns to the functioning of the model as it solves a complex problem, such as the Tower of Hanoi (see Bassok & Novick, Chapter 21). In a series of studies that attempted to isolate cognitive processes in ACT-R with relatively simple tasks, Anderson et al. (2005) associated a posterior parietal brain region with changes in problem representation, a prefrontal brain region with retrieval of task-relevant information, and an area in primary motor cortex with programming of manual responses.

Hummel and Holyoak's LISA model (1997, 2003; see also Doumas & Hummel, Chapter 5, and Holyoak, Chapter 13) is a symbolic-connectionist model that aims to provide a unified account of analogical retrieval, mapping, inference, and relational generalization. LISA solves the relational binding problem by a form of dynamic binding implemented via temporal synchrony. Temporal synchrony is a fundamental property of neural circuits, as evidenced by the rhythmic oscillations evident in raw EEG signals. The use of temporal synchrony as a binding mechanism was first demonstrated via single-cell recording in cat visual cortex (Gray, Engel, Konig, & Singer, 1992), and it has been proposed as a general binding mechanism in the brain (see Singer, 1999). LISA uses temporal synchrony to bind relational representations in working memory. Specifically, propositional structures like *chase* (cat, mouse) are "thought about" in LISA by firing semantic units capturing what it is to chase, and to be a cat, at the same time; and conversely firing units capturing what it is to be chased, and to be a mouse, at the same time, but out of synchrony with *chase* (cat). See Morrison, Doumas, and Richland (2011) for a recent detailed description of LISA's functioning using these types of representations.

While the early papers on LISA remained agnostic regarding the neural substrate of different functions in the model, cognitive neuroscience studies provide a basis for some conjectures (see Fig. 6.5). Based on the previously discussed study by Fuster, Bauer, and Jervey (1985), it appears likely that DLPFC is involved in activating representations for objects and relations that are maintained in posterior cortex (Fig. 6.5b). This process

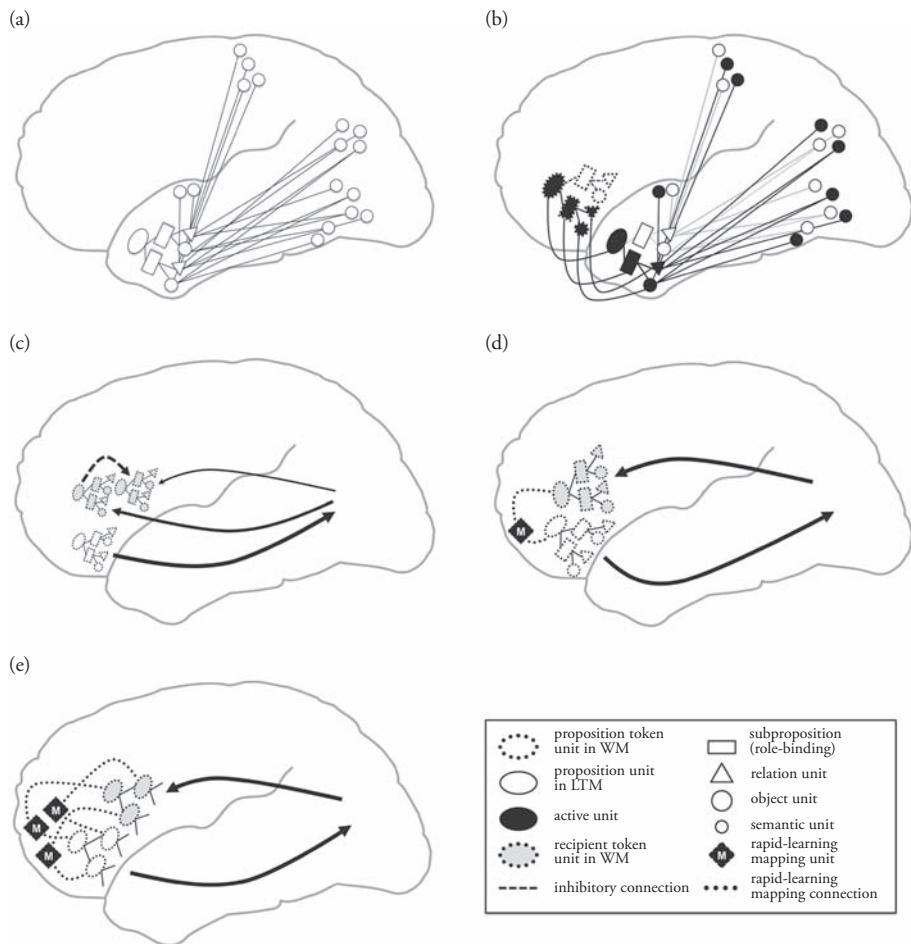


Fig. 6.5 LISA in the brain. (a) Networks responsible for relational representations in long-term memory are likely located in anterior temporal cortex, with their distributed semantic units found in regions responsible for sensation and perception, language, and spatial processing. (b) Thinking about a proposition (e.g., *loves* (Bob, Debbie)) in LISA entails forming dynamic proxy units (Duncan, 2001) in prefrontal cortex (PFC). LISA fires separate subpropositions (e.g., *lover+* Bob) and their connected object (e.g., Bob) and relation (e.g., *loves*) units synchronously. These units pass activation to their static counterparts in long-term memory, effectively bringing long-term memory representations into active working memory (Cowan, 1995; Fuster et al., 1985). Subpropositions for a given proposition alternate being activated in working memory through a process of reciprocal activation and inhibition (i.e., black-and-white units in figure would oscillate out of synchrony with one another). (c) An analog's activation is computed by the rapid-learning mapping units by integrating (over time) the activations of all propositions in that analog. Because of the distributed nature of representations in LISA, more than one potential recipient analog may initially be activated. The probability that a given analog will be retrieved from long-term memory is proportional to its activation divided by the sum of the activations of all analogs. (d) During the pattern of synchronous firing, activation spreads through semantic units in posterior regions of the cortex and gradually activates analogous units in the recipient. Once again dynamic proxy units form in prefrontal cortex. Rapid-learning mapping units in prefrontal cortex (Assad et al., 1998; Cromer, Machon & Miller, 2010) track synchronously active units between the driver and recipient analogs via Hebbian learning and thus learn analogical mappings. Although in this figure only proposition units are shown connected via a rapid-learning mapping unit, all dynamic proxy units in prefrontal cortex (i.e., proposition, subproposition, object, relation) are hypothesized to connect to analogous units via these mapping-unit neurons. (e) Higher order relations (propositions taking propositions as fillers) activate more anterior regions of prefrontal cortex (e.g., Kroger et al., 2002), most likely because of the greater need to track analogical mappings via rapid-learning mapping units.

likely also requires the involvement of IFG to inhibit competing representations (Cho et al., 2010). LISA assumes that propositions are stored in long-term memory as conjunctive codes, perhaps in anterior temporal cortex (Morrison et al.,

2004), with the hippocampus serving to form and later retrieve the conjunctive codes. The process of analogical mapping appears to be dependent on rostralateral prefrontal cortex (Fig. 6.5d; Bunge, Helskog, & Wendelken, 2009; Bunge, Wendelken,

Badre, & Wagner, 2005; Nikitin & Morrison, 2011). The role of the rostralateral prefrontal cortex (RLPFC) in reasoning appears to increase with the complexity of the task-relevant relations (Fig. 6.5e; Kroger et al., 2002). The anterior cingulate is also important for conflict monitoring, particularly in preparation of a decision concerning mapping (Cho et al., 2010; Kroger et al., 2002; Kroger et al., 2008).

The Role of the Frontal Lobes in Human Intelligence

A consistent finding across the range of methodologies employed in cognitive neuroscience is that the frontal lobes play a crucial role in high-level cognitive abilities (Cabeza & Nyberg, 2000; Duncan & Owen, 2000). But while the role of the frontal lobes in intellectual function is fundamental, it is also circumscribed. Intelligence has been typically understood as being of two different forms: crystallized intelligence, including semantic knowledge; and fluid intelligence, which supports abstract reasoning, especially in novel situations. Fluid and crystallized intelligence are differentially affected by aging and brain damage, indicating different neural substrates. Whereas patients with degenerative disease involving the temporal lobes exhibit impaired knowledge of concepts and categories (Krawczyk et al., 2008; Miller, 2007; Morrison et al., 2004), patients with frontal lobe involvement show deficits in problem solving (Holyoak & Kroger, 1995). Frontoparietal circuits are active during fMRI scanning while subjects perform fluid reasoning tasks, such as the Raven's Progressive Matrices (Kroger et al., 2002; Prabhakaran, Smith, Desmond, Glover, & Gabrieli, 1997). It appears that people with a high degree of fluid intelligence further engage regions including the DLPFC and posterior parietal lobule when problems become more complex, whereas individuals with low fluid intelligence scores activate these regions more than those with high fluid intelligence when solving easier problems but do not show this increase when problems become more complex (Perfetti et al., 2009). It is as if moderately difficult problems "saturate" this frontoparietal system in those with low fluid intelligence. High fluid intelligence may be characterized by the ability to effectively engage these circuits.

The comparative anatomy of the frontal lobes, and the fact that they are more highly developed in humans than in other animals, supports the role for the frontal lobes in cognitive abilities that

are most highly developed in humans (Robin & Holyoak, 1995). The posterior part of the frontal lobes includes motor cortical regions, while more anterior, prefrontal regions have more abstract functions. Given the large size of the prefrontal cortex, it seems likely that different subregions have different functions. While there are distinct cytoarchitectonic subregions, these regions are heavily interconnected, suggesting cooperation.

Data from neuroimaging studies often show a great deal of convergence in terms of the regions that are activated in the prefrontal cortex. A number of different tasks involving cognitive control, which includes processes such as resolving response conflict, generation, working-memory manipulation, and categorization, all appear to activate a common set of regions in the lateral prefrontal cortex, dorsal anterior cingulate, and often the premotor cortex (Duncan & Owen, 2000). An increasingly popular method used to define circuits is to assess functional connectivity by measuring the correlation of BOLD signal in different brain regions (Cole & Schneider, 2007). For example, activity in the DLPFC and the anterior cingulate is correlated during task performance, and also when the subject is at rest and not performing any task. These findings indicate that these regions are interconnected and comprise part of a functional network for cognitive control. This network may subserve some process that is common to a diverse set of tasks requiring cognitive control. On the other hand, it may be that this network supports several different functions. If the latter is the case, it may prove difficult to pinpoint the cognitive functions of the prefrontal cortex because these circuits contribute flexibly to cognition.

The cognitive control network comprises a large area of cortex. According to one characterization of the prefrontal cortical control network, there is a hierarchical organization of lateral prefrontal cortex such that the more caudal regions, such as premotor cortex, are involved in response selection based on sensory stimuli, whereas more anterior regions, such as the DLPFC, are necessary when the appropriate response depends on the context. When action depends on retrieved episodic memories, the most anterior regions of the lateral cortex, including the frontal pole, are activated. By this view, the successive layers exert control over each action plan (Koechlin, Ody, & Kouneiher, 2003). In this way, behavior can be modified based on varying levels of complexity that depend on the task demands (Robin & Holyoak, 1995).

In addition to a cognitive control network, other prefrontal regions appear to be activated under other circumstances. Medial and orbital regions appear to be engaged when tasks have emotional or social components (Price, Carmichael, & Drevets, 1996). In addition, the RLPFC, the most anterior part of the prefrontal cortex, appears to have functions distinct from those of lateral prefrontal cortex. A major topic in the cognitive neuroscience of the frontal lobe is the delineation of the contribution of the RLPFC (Bunge et al., 2009; Christoff, Ream, Geddes, & Gabrieli, 2003). As discussed in the hierarchical model of cognitive control, this region becomes important when information retrieved from episodic memory is necessary for forming an action plan. This region is frequently engaged during episodic memory retrieval (Lepage, Ghaffar, Nyberg, & Tulving, 2000), but it is also active while solving reasoning problems that do not have much of a memory retrieval component. According to one view, the RLPFC becomes engaged when the problem requires the integration of multiple relations (Cho et al., 2010; Wendelken & Bunge, 2010); while according to another view, this region becomes engaged when the relations are sufficiently abstract (Christoff, Keramatian, Gordon, Smith, & Madler, 2009). As with the lateral regions, the RLPFC may play multiple roles in cognition. An important direction for future studies is the meta-analysis of neuroimaging data in order to identify commonalities in activation patterns across studies.

Conclusions and Future Directions

Methods in contemporary cognitive neuroscience range from those that have been in use for more than a century (cognitive neuropsychology) to those still undergoing development today. While each technique is better suited to address certain types of questions than others, convergent evidence from multiple methods has been most effective in moving theory forward. Because of the importance of the frontal lobes in complex cognition, their relative accessibility is a boon to researchers using techniques that are limited to the cortical surface. As a more complete understanding of the workings of the prefrontal cortex emerges, perspectives on the nature of thinking will be constrained, or perhaps new perspectives will arise.

Understanding how the brain implements human thinking and reasoning is just in its infancy. The next 10 years promise to be very exciting as the field develops with the use of new methods

and analysis techniques. We believe that the fundamental challenge to this pursuit is to move beyond localist conceptions of brain function toward an understanding of how brain networks develop and operate. Thinking and reasoning are the pinnacles of human cognition and doubtless draw on many different cognitive functions. Understanding how these cognitive functions are harnessed is critical to a fuller understanding of human thought. To get to this point, we believe cognitive neuroscience needs to continue to develop in three core areas.

Cortical Connectivity

Methods for studying when and how brain regions communicate with each other are at the heart of this greater pursuit. Techniques like dynamic causal modeling have provided a way to assess connectivity using fMRI data at a macrodynamic scale; however, we need techniques to study how brain regions communicate at the temporal scale of the timing of neurons. Fortunately, methods such as EEG and MEG and possibly NIRS or EROS provide opportunities to examine real-time temporal dynamics.

Integration of Spatial and Temporal Functional Imaging

While methods like EEG and MEG provide great hope for investigating the temporal dynamics of brain circuits, they have intrinsic limitations for understanding where signals are originating in the brain. Ultimately, researchers interested in understanding brain dynamics will have to develop methods to use spatial localization techniques, such as fMRI, NIRS, or EROS, to target regions of interest, which will help to provide confidence in source localization using EEG and MEG.

Neurocomputational Approaches

Given the complexity of neural systems, we will certainly need principled ways of generating hypotheses about how they function in the service of thinking and reasoning. Computational modeling has been the great friend of thinking and reasoning in the past, helping us to develop and test models at the algorithmic and representation levels of analysis. In the coming years models will need to evolve to achieve realistic neural plausibility and thereby help to make predictions about how neural circuits work together in the service of higher cognition. This will almost certainly involve a merging of different approaches, including symbolic, connectionist, and Bayesian representations and algorithms.

Acknowledgments

The authors would like to thank Keith Holyoak and Slava Nikitin for helpful comments on an earlier draft of the chapter and the American Federation of Aging/Rosalinde and Arthur Gilbert Foundation and the Loyola University Chicago Dean of Arts and Sciences for their generous support (RGM) during the preparation of this manuscript.

Notes

1. Terminology in cognitive neuroscience is frequently rather confusing because of the interdisciplinary origins of the field. In this chapter we use the term “cognitive neuropsychology” to refer to studies of brain-damaged patients, which are frequently based on the logic of single or double dissociations (see Fig. 6.2). Cognitive neuropsychology should not be confused with the field of clinical neuropsychology, which is based on the psychometric appraisal of cognitive function. The cognitive neuropsychology approach is also sometimes referred to as “behavioral neurology.”

2. Frontotemporal lobar degeneration (FTLD) is the newer nomenclature for a syndrome previously referred to as frontotemporal dementia (Miller, 2007). The umbrella of FTLD also frequently includes patients diagnosed with primary progressive aphasia (PPA; Mesulam, 2007). Patients primarily with damage in anterior to dorsolateral frontal cortex are typically referred to as either frontal-variant or behavioral-variant and have symptoms consistent with traditional frontal lobe syndromes (i.e., disinhibition, poor judgment, loss of motivation, executive and working-memory deficits). Patients with damage in anterior temporal cortex, particularly the temporal poles, are frequently referred to as temporal-variant, semantic dementia, or semantic-subtype PPA. Patients diagnosed with FTLD have a range of different postmortem pathologies, including Pick’s disease, cortical basal degeneration and sometimes Alzheimer’s disease.

References

- Anderson, J. R., Albert, M. V., & Fincham, J. M. (2005). Tracing problem solving in real time: fMRI analysis of the subject-paced tower of Hanoi. *Journal of Cognitive Neuroscience*, 17(8), 1261–1274. doi:10.1162/0898929055002427
- Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111(4), 1036–1060. doi:10.1037/0033-295X.111.4.1036
- Ashburner, J., & Friston, K. J. (2000). Voxel-based morphometry—the methods. *NeuroImage*, 11(6 Pt 1), 805–821. doi:10.1006/nimg.2000.0582
- Braine, M. D. S. (1998). Steps toward a mental-predicate logic. In M. D. S. Braine & D. P. O’Brien (Eds.), *Mental logic* (pp. 273–331). Mahwah, NJ: Erlbaum.
- Braver, T. S., Barch, D. M., & Cohen, J. D. (1999). Cognition and control in schizophrenia: A computational model of dopamine and prefrontal function. *Biological Psychiatry*, 46(3), 312–328.
- Brodmann, K. (2006). *Brodmann's localisation in the cerebral cortex the principles of comparative localisation in the cerebral cortex based on the cytoarchitectonics [Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues]* (L. Garey Trans.). New York: Springer.
- Brown, G. G., Perthen, J. E., Liu, T. T., & Buxton, R. B. (2007). A primer on functional magnetic resonance imaging. *Neuropsychology Review*, 17(2), 107–125. doi:10.1007/s11065-007-9028-8
- Bunge, S. A., Helskog, E. H., & Wendelken, C. (2009). Left, but not right, rostral-lateral prefrontal cortex meets a stringent test of the relational integration hypothesis. *NeuroImage*, 46(1), 338–342. doi:10.1016/j.neuroimage.2009.01.064
- Bunge, S. A., Wendelken, C., Badre, D., & Wagner, A. D. (2005). Analogical reasoning and prefrontal cortex: Evidence for separable retrieval and integration mechanisms. *Cerebral Cortex*, 15(3), 239–249. doi:10.1093/cercor/bhh126
- Cabeza, R., & Nyberg, L. (2000). Imaging cognition II: An empirical review of 275 PET and fMRI studies. *Journal of Cognitive Neuroscience*, 12(1), 1–47.
- Canolty, R. T., Edwards, E., Dalal, S. S., Soltani, M., Nagarajan, S. S., Kirsch, H. E.,...Knight, R. T. (2006). High gamma power is phase-locked to theta oscillations in human neocortex. *Science*, 313(5793), 1626–1628. doi:10.1126/science.1128115
- Canolty, R. T., & Knight, R. T. (2010). The functional role of cross-frequency coupling. *Trends in Cognitive Sciences*, 14(11), 506–515. doi:DOI: 10.1016/j.tics.2010.09.001
- Cho, S., Moody, T. D., Fernandino, L., Mumford, J. A., Poldrack, R. A., Cannon, T. D.,...Holyoak, K. J. (2010). Common and dissociable prefrontal loci associated with component mechanisms of analogical reasoning. *Cerebral Cortex*, 20(3), 524–533. doi:10.1093/cercor/bhp121
- Christoff, K., Keramatian, K., Gordon, A. M., Smith, R., & Madler, B. (2009). Prefrontal organization of cognitive control according to levels of abstraction. *Brain Research*, 1286, 94–105. doi:10.1016/j.brainres.2009.05.096
- Christoff, K., Ream, J. M., Geddes, L. P., & Gabrieli, J. D. (2003). Evaluating self-generated information: Anterior prefrontal contributions to human cognition. *Behavioral Neuroscience*, 117(6), 1161–1168. doi:10.1037/0735-7044.117.6.1161
- Ciesielski, K. T., Ahlfors, S. P., Bedrick, E. J., Kerwin, A. A., & Hamalainen, M. S. (2010). Top-down control of MEG alpha-band activity in children performing categorical N-back task. *Neuropsychologia*, 48(12), 3573–3579. doi:10.1016/j.neuropsychologia.2010.08.006
- Cohen, N. J., & Squire, L. R. (1980). Preserved learning and retention of pattern-analyzing skill in amnesia: Dissociation of knowing how and knowing that. *Science*, 210(4466), 207–210.
- Cole, M. W., & Schneider, W. (2007). The cognitive control network: Integrated cortical regions with dissociable functions. *NeuroImage*, 37(1), 343–360. doi:10.1016/j.neuroimage.2007.03.071
- Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user’s guide. *Psychonomic Bulletin and Review*, 12(5), 769–786.
- Conway, A. R., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, 7(12), 547–552.
- Cowan, N. (1995). *Attention and memory: An integrated framework*. New York: Oxford University Press.
- Cromer, J. A., Machon, M., & Miller, E. K. (2010). Rapid association learning in the primate prefrontal cortex in the absence of behavioral reversals. *Journal of Cognitive Neuroscience*, 23, 1823–1828. doi:10.1162/jocn.2010.21555
- Cromer, J. A., Roy, J. E., & Miller, E. K. (2010). Representation of multiple, independent categories in the primate prefrontal

- cortex. *Neuron*, 66(5), 796–807. doi:10.1016/j.neuron.2010.05.005
- Davidson, R. J. (1993). Cerebral asymmetry and emotion: Conceptual and methodological conundrums. *Cognition and Emotion*, 7(1), 115–138. doi:10.1080/02699939308409180
- D'Esposito, M. (2010). Why methods matter in the study of the biological basis of the mind: A behavioral neurologist's perspective. In M. S. Gazzaniga & P. A. Reuter-Lorenz (Eds.), *The cognitive neuroscience of mind: A tribute to Michael S. Gazzaniga* (pp. 203–222). Cambridge, MA: MIT Press.
- Dronkers, N. F., Plaisant, O., Iba-Zizen, M. T., & Cabanis, E. A. (2007). Paul Broca's historic cases: High resolution MR imaging of the brains of Leborgne and Lelong. *Brain*, 130(Pt 5), 1432–1441. doi:10.1093/brain/awm042
- Dumontheil, I., Houlton, R., Christoff, K., & Blakemore, S. J. (2010). Development of relational reasoning during adolescence. *Developmental Science*, 13(6), F15–F24. doi:10.1111/j.1467-7687.2010.01014.x; 10.1111/j.1467-7687.2010.01014.x
- Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nature Review: Neuroscience*, 2(11), 820–829. doi:10.1038/35097575
- Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*, 23(10), 475–483.
- Eslinger, P. J., Blair, C., Wang, J., Lipovsky, B., Realmuto, J., Baker, D.,...Yang, Q. X. (2009). Developmental shifts in fMRI activations during visuospatial relational reasoning. *Brain and Cognition*, 69(1), 1–10. doi:10.1016/j.bandc.2008.04.010
- Feinberg, T. E., & Farah, M. J. (2003). *Behavioral neurology and neuropsychology*. New York: McGraw-Hill Medical Pub. Division.
- Filler, A. (2009). Magnetic resonance neurography and diffusion tensor imaging: Origins, history, and clinical impact of the first 50,000 cases with an assessment of efficacy and utility in a prospective 5000-patient study group. *Neurosurgery*, 65(4 Suppl), A29–A43. doi:10.1227/01.NEU.0000351279.78110.00
- Fischl, B., & Dale, A. M. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences of the United States of America*, 97(20), 11050–11055. doi:10.1073/pnas.200033797
- Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, 19(4), 1273–1302.
- Fuster, J. M., & Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science*, 173(979), 652–654.
- Fuster, J. M., Bauer, R. H., & Jervey, J. P. (1985). Functional interactions between inferotemporal and prefrontal cortex in a cognitive task. *Brain Research*, 330(2), 299–307.
- Galambos, R., & Sheatz, G. C. (1962). An electroencephalograph study of classical conditioning. *The American Journal of Physiology*, 203, 173–184.
- Glover, G. H. (1999). Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage*, 9(4), 416–429.
- Goel, V. (2007). Anatomy of deductive reasoning. *Trends in Cognitive Sciences*, 11(10), 435–441. doi:10.1016/j.tics.2007.09.003
- Goel, V., Buchel, C., Frith, C., & Dolan, R. J. (2000). Dissociation of mechanisms underlying syllogistic reasoning. *NeuroImage*, 12(5), 504–514. doi:10.1006/nimg.2000.0636
- Goel, V., & Dolan, R. J. (2003). Explaining modulation of reasoning by belief. *Cognition*, 87, B11–B22.
- Goel, V., & Dolan, R. J. (2004). Differential involvement of left prefrontal cortex in inductive and deductive reasoning. *Cognition*, 93(3), B109–B121. doi:10.1016/j.cognition.2004.03.001
- Gratton, G., Fabiani, M., Corballis, P. M., Hood, D. C., Goodman-Wood, M. R., Hirsch, J.,...Gratton, E. (1997). Fast and localized event-related optical signals (EROS) in the human occipital cortex: Comparisons with the visual evoked potential and fMRI. *NeuroImage*, 6(3), 168–180. doi:DOI: 10.1006/nimg.1997.0298
- Gray, C. M., Engel, A. K., Konig, P., & Singer, W. (1992). Synchronization of oscillatory neuronal responses in cat striate cortex: Temporal properties. *Visual Neuroscience*, 8(4), 337–347.
- Hansen, P. C., Kringlebach, M. L., & Salmelin, R. (2010). *MEG: An introduction to methods*. New York: Oxford University Press.
- Hinton, E. C., Dymond, S., von Hecker, U., & Evans, C. J. (2010). Neural correlates of relational reasoning and the symbolic distance effect: Involvement of parietal cortex. *Neuroscience*, 168(1), 138–148. doi:10.1016/j.neuroscience.2010.03.052
- Holyoak, K. J., & Kroger, J. K. (1995). Forms of reasoning: Insight into prefrontal functions? *Annals of the New York Academy of Sciences USA*, 769, 253–263.
- Huey, E. D., Goveia, E. N., Paviol, S., Pardini, M., Krueger, F., Zamboni, G.,...Grafman, J. (2009). Executive dysfunction in frontotemporal dementia and corticobasal syndrome. *Neurology*, 72(5), 453–459. doi:10.1212/01.wnl.0000341781.39164.26
- Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104(3), 427–466. doi:10.1037/0033-295X.104.3.427
- Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, 110(2), 220–264.
- Humphrey, D. R., & Schmidt, E. M. (1990). Extracellular single-unit recording methods. In A. A. Boulton, G. B. Baker, & C. H. Vanderwolf (Eds.), *Neurophysiological techniques: Applications to neural systems* (pp. 1–64). Clifton, NJ: Springer. doi:10.1385/0-89603-185-3:1
- Jung-Beeman, M., Bowden, E. M., Haberman, J., Fryniare, J. L., Arambel-Liu, S., Greenblatt, R.,...Kounios, J. (2004). Neural activity when people solve verbal problems with insight. *PLoS Biology*, 2(4), E97. doi:10.1371/journal.pbio.0020097
- Kean, M. (1977). The linguistic interpretation of aphasic syndromes: Agrammatism in Broca's aphasia, an example. *Cognition*, 5(1), 9–46. doi:10.1016/0010-0277(77)90015-4
- Kiebel, S. J., Garrido, M. I., Moran, R., Chen, C. C., & Friston, K. J. (2009). Dynamic causal modeling for EEG and MEG. *Human Brain Mapping*, 30(6), 1866–1876. doi:10.1002/hbm.20775
- Knauff, M., Fangmeier, T., Ruff, C. C., & Johnson-Laird, P. N. (2003). Reasoning, models, and images: Behavioral measures and cortical activity. *Journal of Cognitive Neuroscience*, 15(4), 559–573. doi:10.1162/089892903321662949
- Knauff, M., & Johnson-Laird, P. N. (2002). Visual imagery can impede reasoning. *Memory and Cognition*, 30(3), 363–371.
- Knight, R. T., & Graboweczyk, M. (1995). Escape from linear time: Prefrontal cortex and conscious experience. In

- M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 1357–1371). Cambridge, MA: MIT Press.
- Knight, R. T., Scabini, D., & Woods, D. L. (1989). Prefrontal cortex gating of auditory transmission in humans. *Brain Research*, 504(2), 338–342.
- Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science*, 273(5280), 1399–1402.
- Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science*, 302(5648), 1181–1185. doi:10.1126/science.1088545
- Kounios, J., Frymiare, J. L., Bowden, E. M., Fleck, J. I., Subramaniam, K., Parrish, T. B., & Jung-Beeman, M. (2006). The prepared mind: Neural activity prior to problem presentation predicts subsequent solution by sudden insight. *Psychological Science*, 17(10), 882–890. doi:10.1111/j.1467-9280.2006.01798.x
- Krawczyk, D. C., Morrison, R. G., Viskontas, I., Holyoak, K. J., Chow, T. W., Mendez, M. F.,... Knowlton, B. J. (2008). Distraction during relational reasoning: The role of prefrontal cortex in interference control. *Neuropsychologia*, 46(7), 2020–2032. doi:10.1016/j.neuropsychologia.2008.02.001
- Kroger, J. K., Nystrom, L. E., Cohen, J. D., & Johnson-Laird, P. N. (2008). Distinct neural substrates for deductive and mathematical processing. *Brain Research*, 1243, 86–103. doi:10.1016/j.brainres.2008.07.128
- Kroger, J. K., Sabb, F. W., Fales, C. L., Bookheimer, S. Y., Cohen, M. S., & Holyoak, K. J. (2002). Recruitment of anterior dorsolateral prefrontal cortex in human reasoning: A parametric study of relational complexity. *Cerebral Cortex*, 12(5), 477–485.
- Lee, D. A. (1981). Paul Broca and the history of aphasia: Roland P. Mackay award essay, 1980. *Neurology*, 31(5), 600–602.
- Lepage, M., Ghaffar, O., Nyberg, L., & Tulving, E. (2000). Prefrontal cortex and episodic memory retrieval mode. *Proceedings of the National Academy of Sciences of the USA*, 97(1), 506–511.
- Luck, S. J. (2005). *An introduction to the event-related potential technique*. Cambridge, MA: MIT Press.
- Luck, S. J., & Kappenman, E. (Eds.) (2012). *Oxford handbook of event-related potential components*. New York: Oxford University Press.
- Luo, Q., Perry, C., Peng, D., Jin, Z., Xu, D., Ding, G., & Xu, S. (2003). The neural substrate of analogical reasoning: An fMRI study. *Brain Research: Cognitive Brain Research*, 17(3), 527–534.
- Mesulam, M. M. (2007). Primary progressive aphasia: A 25-year retrospective. *Alzheimer Disease and Associated Disorders*, 21(4), S8–S11. doi:10.1097/WAD.0b013e31815bf7e1
- Miller, B. L. (2007). Frontotemporal dementia and semantic dementia: Anatomic variations on the same disease or distinctive entities? *Alzheimer Disease and Associated Disorders*, 21(4), S19–S22. doi:10.1097/WAD.0b013e31815c0f7a
- Miller, K. J., denNijs, M., Shenoy, P., Miller, J. W., Rao, R. P. N., & Ojemann, J. G. (2007). Real-time functional brain mapping using electrocorticography. *NeuroImage*, 37(2), 504–507. doi:DOI: 10.1016/j.neuroimage.2007.05.029
- Monti, M. M., Osherson, D. N., Martinez, M. J., & Parsons, L. M. (2007). Functional neuroanatomy of deductive inference: A language-independent distributed network. *Neuroimage*, 37, 1005–1016. doi:10.1016/j.neuroimage.2007.04.069
- Monti, M. M., Parsons, L. M., & Osherson, D. N. (2009). The boundaries of language and thought in deductive inference. *Proceedings of the National Academy of Sciences of the USA*, 106(30), 12554–12559. doi:10.1073/pnas.0902422106
- Morrison, R. G. (2005). Thinking in working memory. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning*. (pp. 457–473). New York: Cambridge University Press.
- Morrison, R. G., Doumas, L. A. A., & Richland, L. E. (2011). A computational account of children's analogical reasoning: Balancing inhibitory control in working memory and relational representation. *Developmental Science*, 14, 516–529. doi:10.1111/j.1467-7687.2010.00999.x
- Morrison, R. G., Krawczyk, D. C., Holyoak, K. J., Hummel, J. E., Chow, T. W., Miller, B. L., & Knowlton, B. J. (2004). A neurocomputational model of analogical reasoning and its breakdown in frontotemporal lobar degeneration. *Journal of Cognitive Neuroscience*, 16(2), 260–271. doi:10.1162/08989290432298453
- Nikitin, S., & Morrison, R. G. (2011). *Analogical reasoning in human prefrontal cortex: An event-related potential approach*. Paper presented at the Cognitive Neuroscience Society Annual Meeting, San Francisco, CA.
- Noveck, I. A., Goel, V., & Smith, K. W. (2004). The neural basis of conditional reasoning with arbitrary content. *Cortex*, 40(4–5), 613–622.
- Okubo, Y., Suhara, T., Suzuki, K., Kobayashi, K., Inoue, O., Terasaki, O.,... Toru, M. (1997). Decreased prefrontal dopamine D1 receptors in schizophrenia revealed by PET. *Nature*, 385, 634–636. doi:10.1038/385634a0
- O'Reilly, R. C. (2006). Biologically based computational models of high-level cognition. *Science*, 314(5796), 91–94. doi:10.1126/science.1127242
- Pascual-Leone, A., Bartres-Faz, D., & Keenan, J. P. (1999). Transcranial magnetic stimulation: Studying the brain-behaviour relationship by induction of 'virtual lesions'. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 354(1387), 1229–1238. doi:10.1098/rstb.1999.0476
- Perfetti, B., Saggino, A., Ferretti, A., Caulo, M., Romani, G. L., & Onofrj, M. (2009). Differential patterns of cortical activation as a function of fluid reasoning complexity. *Human Brain Mapping*, 30(2), 497–510. doi:10.1002/hbm.20519
- Prabhakaran, V., Smith, J. A., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. (1997). Neural substrates of fluid reasoning: An fMRI study of neocortical activation during performance of the raven's progressive matrices test. *Cognitive Psychology*, 33(1), 43–63. doi:10.1006/cogp.1997.0659
- Prado, J., & Noveck, I. A. (2007). Overcoming perceptual features in logical reasoning: A parametric functional magnetic resonance study. *Journal of Cognitive Neuroscience*, 19, 642–657. doi:10.1162/jocn.2007.19.4.642
- Prado, J., Van Der Henst, J. B., & Noveck, I. A. (2010). Recomposing a fragmented literature: How conditional and relational arguments engage different neural systems for deductive reasoning. *NeuroImage*, 51(3), 1213–1221. doi:10.1016/j.neuroimage.2010.03.026
- Price, J. L., Carmichael, S. T., & Drevets, W. C. (1996). Networks related to the orbital and medial prefrontal cortex: a substrate for emotional behavior? *Progress in Brain Research*, 107, 523–536.
- Raichle, M. E. (1983). Positron emission tomography. *Annual Review of Neuroscience*, 6, 249–267. doi:10.1146/annurev.ne.06.030183.001341

- Raichle, M. E. (1994). Visualizing the mind. *Scientific American*, 270(4), 58–64.
- Raichle, M. E., & Mintun, M. A. (2006). Brain work and brain imaging. *Annual Review of Neuroscience*, 29, 449–476. doi:10.1146/annurev.neuro.29.051605.112819
- Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. Cambridge, MA: MIT Press.
- Robin, N., & Holyoak, K. J. (1995). Relational complexity and the functions of prefrontal cortex. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 987–997). Cambridge, MA: MIT Press.
- Rogalski, E. J., Murphy, C. M., deToledo-Morrell, L., Shah, R. C., Moseley, M. E., Bammer, R., & Stebbins, G. T. (2009). Changes in parahippocampal white matter integrity in amnestic mild cognitive impairment: A diffusion tensor imaging study. *Behavioral Neurology*, 21(1), 51–61. doi:10.3233/BEN-2009-0235
- Rosen, H. J., Gorno-Tempini, M. L., Goldman, W. P., Perry, R. J., Schuff, N., Weiner, M.,...Miller, B. L. (2002). Patterns of brain atrophy in frontotemporal dementia and semantic dementia. *Neurology*, 58(2), 198–208.
- Sauseng, P., & Klimesch, W. (2008). What does phase information of oscillatory brain activity tell us about cognitive processes? *Neuroscience and Biobehavioral Reviews*, 32(5), 1001–1013. doi:10.1016/j.neubiorev.2008.03.014
- Singer, W. (1999). Neuronal synchrony: A versatile code for the definition of relations? *Neuron*, 24(1), 49–65, 111–125.
- Sonty, S. P., Mesulam, M. M., Weintraub, S., Johnson, N. A., Parrish, T. B., & Gitelman, D. R. (2007). Altered effective connectivity within the language network in primary progressive aphasia. *Journal of Neuroscience*, 27(6), 1334–1345. doi:10.1523/JNEUROSCI.4127–06.2007
- Tsujii, T., Masuda, S., Akiyama, T., & Watanabe, S. (2010). The role of inferior frontal cortex in belief-bias reasoning: An rTMS study. *Neuropsychologia*, 48(7), 2005–2008. doi:10.1016/j.neuropsychologia.2010.03.021
- Tsujii, T., & Watanabe, S. (2010). Neural correlates of belief-bias reasoning under time pressure: A near-infrared spectroscopy study. *NeuroImage*, 50(3), 1320–1326. doi:10.1016/j.neuroimage.2010.01.026
- Villemagne, V. L., Pike, K. E., Chetelat, G., Ellis, K. A., Mulligan, R. S., Bourgeat, P.,...Rowe, C. C. (2011). Longitudinal assessment of abeta and cognition in aging and alzheimer disease. *Annals of Neurology*, 69(1), 181–192. doi:10.1002/ana.22248; 10.1002/ana.22248
- Villringer, A., & Chance, B. (1997). Non-invasive optical spectroscopy and imaging of human brain function. *Trends in Neurosciences*, 20(10), 435–442.
- Viskontas, I. V., Morrison, R. G., Holyoak, K. J., Hummel, J. E., & Knowlton, B. J. (2004). Relational integration, inhibition, and analogical reasoning in older adults. *Psychology and Aging*, 19(4), 581–591. doi:10.1037/0882-7974.19.4.581
- Vogel, E. K., McCollough, A. W., & Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature*, 438(7067), 500–503. doi:10.1038/nature04171
- Waltz, J. A., Knowlton, B. J., Holyoak, K. J., Boone, K. B., Mishkin, F. S., de Menezes Santos, M.,...Miller, B. L. (1999). A system for relational reasoning in human prefrontal cortex. *Psychological Science*, 10(2), 119–125. doi:10.1111/1467-9280.00118
- Wassermann, E. M., & Lisanby, S. H. (2001). Therapeutic application of repetitive transcranial magnetic stimulation: A review. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, 112(8), 1367–1377.
- Wendelken, C., & Bunge, S. A. (2010). Transitive inference: Distinct contributions of rostral-lateral prefrontal cortex and the hippocampus. *Journal of Cognitive Neuroscience*, 22(5), 837–847. doi:10.1162/jocn.2009.21226
- Zamboni, G., Huey, E. D., Krueger, F., Nichelli, P. F., & Grafman, J. (2008). Apathy and disinhibition in frontotemporal dementia: Insights into their neural correlates. *Neurology*, 71(10), 736–742. doi:10.1212/01.wnl.0000324920.96835.95

Mental Function as Genetic Expression: Emerging Insights From Cognitive Neurogenetics

Adam E. Green and Kevin N. Dunbar

Abstract

Following the decade (or two) of the brain, a new effort is underway to integrate insights about the biology of mental function that have been gained at parallel levels of description, in molecular genetics, cognitive neuroscience, and psychology. Integrative cognitive neurogenetic research promises new contributions to our understanding of how genes affect the mind by shaping the brain. These contributions include research into psychological functions that do not readily lend themselves to animal models. Despite the complex pathway from genetic variation to changes in psychological functions and behavior—the effects of any one gene depend on its interaction with other genes and with the environment—molecular-genetic data have the potential to inform psychological function in unique ways. We review work on working memory, attention, long-term memory, and language that illustrates this emerging potential.

Key Words: cognitive neurogenetic, intermediate phenotype, imaging genetic, cognitive neuroscience, molecular genetics

Introduction

As the focus of psychological inquiry has shifted to physical characteristics of the human brain, especially with the advent of direct, *in vivo* assays of brain function, some of the spirit has gone out of the long debate between dualism and materialism. And if we accept the materialist premise that the physical brain constitutes the mind, a consequence is that the genome, which directs the brain's physical construction, must shape mental function. An established behavioral-genetic literature has demonstrated that genes strongly influence human cognition, emotion, and subjective experience, often in ways that depend critically on the environment (Caspi et al., 2003; Turkheimer, Haley, Waldron, D'Onofrio, & Gottesman, 2003). Recently, a new research area has taken root at the intersection of psychology, neuroscience, and molecular genetics (Goldberg & Weinberger, 2004; Green et al., 2008; Meyer-Lindenberg & Weinberger, 2006). A major

motivation for this "cognitive neurogenetic" research has been to relate genetic variation to brain function, especially as a window into genetic susceptibility to psychiatric disorders (Goldberg & Weinberger, 2004; Meyer-Lindenberg & Weinberger, 2006; Tan, Callicott, & Weinberger, 2008; see Bachman & Cannon, Chapter 34). While we address some of this foundational work, the review presented in this chapter primarily emphasizes the potential of molecular genetics to inform our understanding of mind-brain relationships more generally in healthy populations. We focus on the role of the cognitive neurogenetic approach in helping to explain psychological functions that we all have in common as well as differences in mental efficacy between individuals.

This chapter is intended as a psychologically oriented primer rather than an exhaustive review. Our goal is to provide a reference for recent developments in cognitive neurogenetics as a starting point for those with a primary interest in psychology to learn more

Table 7.1. Some psychological phenotypes, associated brain-based intermediate phenotype, and putative genes of influence

Psychological Function	Intermediate Phenotype	Implicated Genes
Working memory	DLPFC efficiency and functional connectivity	Dopamine-related: <i>COMT, DAT1</i>
Attention		
Alerting	Thalamic activity	Noradrenaline-related: <i>ADRA2A, NET</i>
Orienting	Parietal activity	Acetylcholine-related: <i>CHNA4, CHRNA7</i>
Executive control	Anterior cingulate activity	Dopamine-related: <i>COMT, DAT1, MAO-A</i>
Long-term memory (encoding and retrieval)	Hippocampal activity Hippocampal and parahippocampal volume Hippocampal N-acetyl-aspartate	Synaptic plasticity-related: <i>BDNF, Prion</i> Dopamine-related: <i>COMT</i> Metabotropic signaling cascade-related: <i>ADCY8, CAMK2G, GRM3, PRKACG</i>
Language	White matter volume and connectivity	Neurotrophic factor-related: <i>PLXNB3, BDNF</i>

about how genetic data are being incorporated into each of several key research areas in psychology. We briefly introduce conceptual issues and then review the extant literature, dividing our review into four areas of experimental psychology: working memory, attention, long-term memory, and language. For quick reference, Table 7.1 lists some genes and associated brain characteristics relevant to the cognitive neurogenetics of these areas of psychology. We take a decidedly cognitive approach to divvying up mental function, and it is worth noting that this is a chapter concerned with relatively modular cognitive operations in a book with the broader charter of thinking and reasoning. While cognitive neurogenetics aspires toward more integrated phenotypes, extant work has focused on specific cognitive operations. Nonetheless, cognitive operations like memory, attention, and language are largely the constituents of thinking and reasoning, so a review of this literature can inform the central questions this volume addresses. In the discussion we assess the logic, benefits, and obstacles that attend cognitive neurogenetic inquiry.

Back to Bases

Before we begin our review, it may be useful to discuss in general terms the sort of molecular-biological

sequence of events that can lead from genes to cognition. Human beings are made up of cells that have a nucleus and within the nucleus there are chromosomes, which contain genes that are the basic units of inheritance (human beings have 23 pairs of chromosomes, which contain roughly 30,000 genes in total). Each gene consists of a sequence of DNA on a chromosome. The process of expressing the information encoded in gene sequences begins with transcription of DNA into RNA, and then continues by means of complex cellular machinery to produce proteins that sustain cells and direct them to perform particular functions. This is what is often referred to as the central dogma of molecular biology. While there are many nuances concerning what exactly constitutes a gene and how genes are transcribed, the fundamental property of genes that is relevant for cognitive neurogenetics is that genes are made from subunits called *nucleotides*. Each nucleotide is made of a sugar and phosphate and a chemical base. There are four different bases—adenine (A), thymine (T), guanine (G), and cytosine (C)—that determine the functional identity of the nucleotide in which they occur. These nucleotides are the fundamental structural and coding elements of DNA. The DNA itself consists of a pair of strands that wind around each

other in the form of a double helix. Because of the charges on the bases and the structure of the helix, only an A on one strand of the helix can match to a T on the other strand, and only a G on one strand can match to a C on the other strand.

A key distinction in genetics research is between the phenotype, which is the observable behavior, and the genotype, which is the underlying genetic coding. As molecular-genetics research has progressed toward mapping entire sequences of nucleotides on a gene, it has been discovered that only some of the DNA is expressed. Indeed, the actual functions of most genetic sequences remain unknown. Furthermore, individuals can have different sequences of nucleotides yet have similar phenotypes. Nevertheless, between-human variations in the sequence of nucleotides within the DNA of a gene can lead to differences in protein transcription that can ultimately have bearing on the complex machinery of the mind. These variations in nucleotide sequence are called polymorphisms. Single nucleotide polymorphisms (SNPs) involve variation of just one nucleotide in a sequence and are relatively common (e.g., an A might be substituted for a T, or a G for a C). Polymorphic variations that have been linked to variation in cognitive phenotypes are a major target of investigation in cognitive neurogenetics. This work will be the primary focus of our review.

Molecular Genetics in Cognitive Neuroscience

To what extent can molecular-genetic data further the broad goal of cognitive neuroscience to understand the mind mechanistically? There are at least three potential benefits of incorporating such data into cognitive neuroscience studies: (1) to tell us more about the brain by providing insight into molecular-biological machineries that contribute to the building of neural systems and implementation of cognitive function; (2) To tell us more about the mind, by providing a new source of explanatory variance (genetic variation) to test hypotheses about psychological mechanisms (e.g., to test whether putatively separate psychological mechanisms are associated with distinct sets of genes); and (3) to tell us more about sources of individual differences by characterizing genetic variables, which are less susceptible to confounding factors (e.g., motivation, prior experience/knowledge, cognitive and emotional state) that are often difficult to remove from standard cognitive neuroscience studies (see Fig. 7.1).

INTERMEDIATE PHENOTYPES

Cognitive neurogenetics takes an *intermediate phenotype* approach (Tan et al., 2008), in that it seeks to link observable psychological phenotypes to polymorphic loci in the genome by way of neural phenotypes (intermediate between genotype and psychological function) that are detectable through brain-based investigation. Intermediate phenotypes are the same neural structures and functions that are the typical object of cognitive neuroscience studies (Goldberg & Weinberger, 2004; Meyer-Lindenberg & Weinberger, 2006). Critically, intermediate phenotypes may be more amenable to genetic investigation than behavioral phenotypes (Goldberg & Weinberger, 2004; Tan, Callicott et al., 2008), the key idea being that they are physiologically more proximate to gene expression than are overt behaviors (Blackwood et al., 2001; Goldberg & Weinberger, 2004; Meyer-Lindenberg & Weinberger, 2006; Tan, Callicott et al., 2008). In contrast, behavior is a more distant, final common pathway—and hence presumably more variable as an outcome measure (Caspi & Moffitt, 2006; Meyer-Lindenberg & Weinberger, 2006; Tan, Callicott et al., 2008).

INTERPRETIVE HAZARDS

There is considerable complexity at each link in the chain of associations from single polymorphism to brain activity to psychological function. Appropriate interpretation requires cognizance of some key conceptual issues, including pleiotropy, polygenicity, and false-positive genetic association. Many of the genes that are relevant to psychological function are *pleiotropic*, that is, they contribute to many functions (Kovas & Plomin, 2006). For example, a gene that influences a general feature of brain structure is likely to influence multiple brain functions that depend on this feature (e.g., the brain-derived neurotrophic factor [*BDNF*] gene is thought to influence the development of white matter connections throughout the brain). Adding to the complexity many psychological functions are influenced by multiple genes and/or multiple variations within a gene, a phenomenon known as *polygenicity* (Fisher, 2006). Furthermore, genes can exert considerable influence on each other (Fisher, 2006). For example, as we will see, there is evidence that human working memory function is shaped by several interacting variations in genes that influence dopamine signaling. Thus, multiple complementary and convergently focused studies are likely to be required to fully understand causal gene-to-phenotype (and gene-to-intermediate phenotype) relationships in detail.

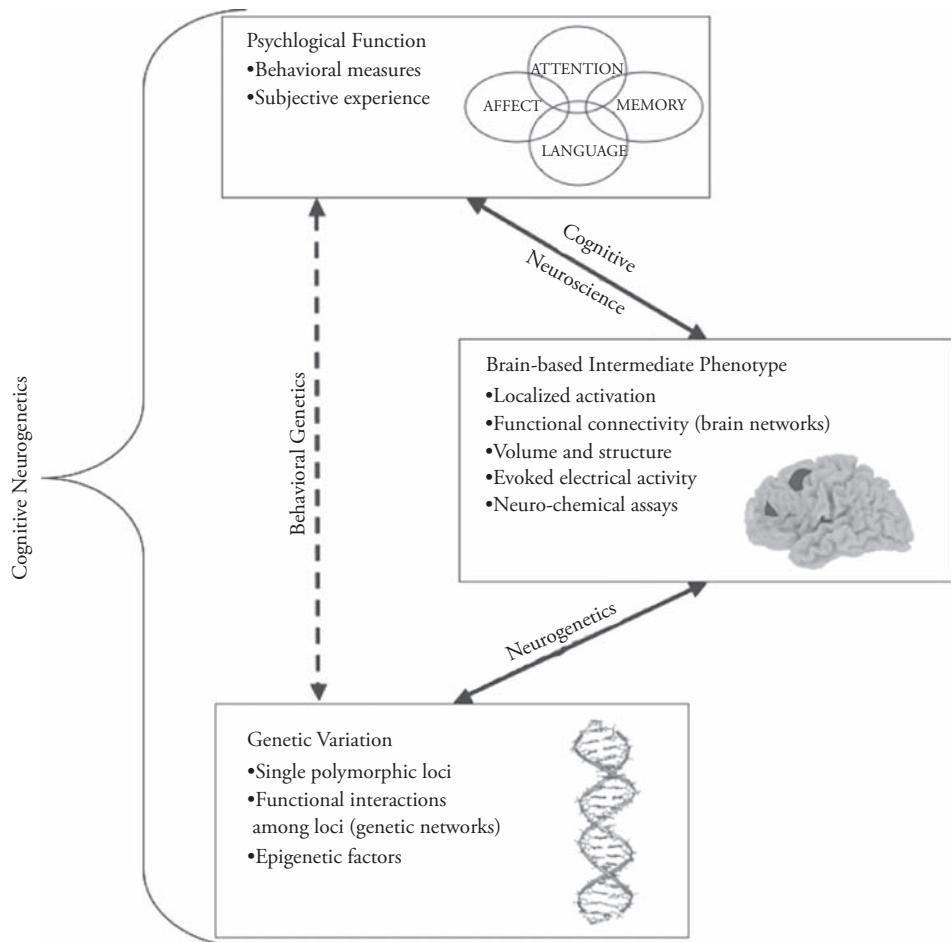


Fig. 7.1 Integrating cognitive neuroscience and molecular genetics through the intermediate phenotype approach of cognitive neurogenetics. As applied to cognitive neuroscience, cognitive neurogenetics seeks to identify links between psychological functions and points of variation in the genome via intermediate neural characteristics that (1) are associated with the psychological function and (2) vary as a function of the genetic variation. This cognitive neurogenetic approach incorporates standard cognitive neuroscience methods to characterize psychological functions in terms of their underlying neural substrates (intermediate phenotypes), and genetic analysis to determine whether specified genetic variables are associated with these intermediate phenotypes. This approach also typically involves behavioral-genetic analysis in order to test whether behavioral measures of the psychological function of interest are influenced by the specified genetic variable. A relationship between genetic variation and an intermediate phenotype, even in the absence of a brain-behavior association can be informative; this sort of finding can identify neural effects of genetically determined differences that would not be observed in standard cognitive neuroscience.

From a methodological perspective, an issue requiring caution in the study of genetic associations is the potential for “false positive” findings (Type I errors) owing to the multitude of genes (and polymorphisms) that can be investigated for a given psychological function. This issue is potentially compounded for cognitive neurogenetics research because neuroimaging datasets involve sampling from multiple points in the brain, heightening the potential risks of multiple comparisons and increasing the necessity of statistical rigor if true gene–brain-cognition associations are to be identified. Illustrating the importance of interpretive

caution in genetic association is the fact that early studies of behavioral and clinical phenotypes found the proportion of phenotypic variance accounted for by a single genetic variant was as much as 10%–20%, whereas subsequent meta-analyses suggested that the true figure is typically closer to 0.1%–1% (Flint & Munafò, 2007). Importantly, strategies such as choosing the genes and neuroanatomical regions of interest *a priori* focusing on genetic variants with a known functional significance, using rigorous statistical criteria, replication, and meta-analyses can reduce the risk of making Type I errors. The intermediate

phenotype approach, when constrained by carefully targeted anatomical hypotheses, has the potential to reduce the risk of Type I error (as compared to the use of behavioral or clinical phenotypes; Meyer-Lindenberg et al., 2008) but cannot eliminate this risk. A point of particular emphasis for cognitive neurogenetics studies should be increased statistical power through large sample sizes.

EPIGENETIC INFLUENCES

Points of variation within DNA have been the primary targets of cognitive neurogenetic investigation, and this emphasis is reflected in our review. However, in addition to this research, exciting molecular-genetic data have emerged, which concern variations not in the genome itself, but in the chromosomal structures that scaffold DNA and can influence its transcription. Such “epigenetic” variations, while not encoded within genes, do exhibit heritability (see Qiu, 2006, for a brief overview) and have consequences for genetic expression in development, protection against disease, and psychological functions (Masterpasqua, 2009; McGowan & Szyf, 2010). Intriguingly, epigenetic research promises insights into mechanisms by which elements of our external environment—physical, social, psychological, and otherwise—produce changes in gene expression. Extant epigenetic research has had a largely clinical focus. While our focus here is primarily nonclinical, an introduction to the concepts and promise of this emerging area is worthwhile because the use of epigenetic data in cognitive neurogenetics is likely to expand.

Epigenetic effects depend on the way that DNA is bound to the chromosomes that carry it. Chromosomes are largely composed of protein units called *histones*, around which strands of DNA are wrapped in convoluted loops. In aggregate, histones and their bound DNA strands are called *chromatin*. Each histone has a tail, which protrudes out from the histone protein. These tails can be bound to a methyl group (methylation), a phosphate group (phosphorylation), or an acetyl group (acetylation). Methylation, phosphorylation, and acetylation alter chromatin structure, thereby influencing the availability of bound strands of DNA to be transcribed to RNA. These alterations modify protein production and, at least in some cases, are brought on by environmental influences.

One of the clearest examples of epigenetic processes at work relates to the question of why identical twins become different over time. Given that these twins are extremely alike at the level of

genotype, how is it that they should differ, sometimes greatly, in phenotype? Fraga and colleagues (2005) investigated the methylation and acetylation of histones in monozygotic twins. These researchers discovered little difference in young twins, but in older twins found large methylation and acetylation differences, which were shown to affect genetic expression. Similar studies using animal models have obtained similar results that indicate that environmental factors as diverse as maternal bonding, diet, smoking, and social interactions can all change the expression of genes in an epigenetic manner (see Relton & Smith, 2010, for a review). Work has been undertaken to draw links between epigenetic variables and neural intermediate phenotypes related to cognition, primarily in clinical populations (Canli et al., 2006; Dulac, 2010; Graff & Mansuy, 2008, 2009; Riccio, 2010). As cognitive neurogenetic models become more complete, it will also be important to characterize interactions between genetic and epigenetic variables in the genesis of healthy cognitive function.

Complex Cognition and Working Memory

The emergence of cognitive neuroscience over the past two decades has enabled novel insights into the most advanced reaches of human mental function (see Morrison & Knowlton, Chapter 6). Data from cognitive neuroscience studies, combined with data from animal research and lesions studies, have reliably identified prefrontal cortex (PFC) as neural substrate for the most complex, “high-level” cognitive operations (Green, Kraemer, Fugelsang, Gray, & Dunbar, 2010; Green, Kraemer, Fugelsang, Gray, & Dunbar, in press; Kane & Engle, 2002; Ramnani & Owen, 2004). Brain networks within PFC—operating in connection with other regions—mediate component demands of complex cognitive tasks. A strong heritability of prefrontal gray matter volume (~80%), and a strong predictive relationship between prefrontal gray matter and general intelligence, indicate that there are genetic influences on PFC-mediated cognition (Bouchard & McGue, 1981; Gray & Thompson, 2004; Thompson et al., 2001). Indeed, many of the most complex prefrontally mediated cognitive traits, including IQ (Allen et al., 2005; Dick et al., 2007; Le Hellard et al., 2009), executive function (Apud et al., 2007; Barnett, Heron, Goldman, Jones, & Xu, 2009; Bishop, Fossella, Croucher, & Duncan, 2008; Egan et al., 2001; Joober et al., 2002; Malhotra et al., 2002; Pezawas et al., 2004; Rybakowski, Borkowska, Skibinska, & Hauser, 2006;

Tan, Nicodemus et al., 2008; Tsai et al., 2008; Waldman et al., 2006), and even creativity (Dunbar, 2008; Reuter, Roth, Holte, & Hennig, 2006), have been linked to specific genetic polymorphisms. This evidence for genetic influence, combined with strong brain-imaging characterization of brain networks underpinning PFC-mediated tasks, suggests that the PFC may be a suitable target for the application of molecular-genetic methods to cognitive neuroscience. Indeed, a growing body of findings has begun to inform mechanistic understanding of molecular-genetic influences on brain networks that mediate complex cognition.

Working memory is a well-characterized form of complex cognition that appears to be central to a range of higher cognitive functions. Working memory refers to the ability to maintain and manipulate information in short-term memory while resisting interference (Baddeley, 2003). It is related to a number of high-level capacities such as reasoning, and its neural substrate overlaps that of intelligence (Gray, Chabris, & Braver, 2003; Just & Carpenter, 1992). Networks of brain areas, including dorsolateral prefrontal cortex (DLPFC) and hippocampus, have been reliably associated with visual and verbal working memory systems, and several tasks have been developed to study individual differences in working memory capacity (Nystrom et al., 2000; Smith & Jonides, 1998; Turner & Engle, 1989). Working memory capacity, and patterns of underlying brain activity, show strong heritability (Koten et al., 2009). Molecular-genetic approaches have focused on genes that influence the dopamine system. This focus stems from a rich clinical literature, especially work in schizophrenia, which has implicated disordered dopaminergic function in pathological impairment of working memory (Cools & Robbins, 2004; Goldberg et al., 2003; Meyer-Lindenberg & Weinberger, 2006; Weinberger et al., 2001; see Bachman & Cannon, Chapter 34).

A dopamine-system-related gene that has been frequently studied for its possible effects on working memory is the catechol-*O*-methyltransferase (*COMT*) gene. *COMT* encodes the COMT enzyme, which catabolically terminates the activity of dopamine in all dopaminergic synapses. In PFC, COMT enzymatic activity is thought to be particularly important for determining dopamine availability because the dopamine reuptake transporter protein, which helps to clear dopamine from the synaptic cleft, shows sparse expression in PFC relative to striatal areas (Lewis et al., 2001; Matsumoto et al.,

2003). Much investigation has therefore targeted a common polymorphism that is known to influence the enzymatic activity of COMT. An evolutionarily recent G-to-A mutation, causing a valine (Val)-to-methionine (Met) substitution at codon 158 of the *COMT* gene on chromosome 22q11, results in relatively reduced COMT enzyme activity. Thus, Met allele carriers have relatively low COMT enzyme activity and have higher dopamine availability as a result, whereas Val allele carriers have relatively increased COMT activity and lower dopamine availability (Lotta et al., 1995).

Some insight into the effects of this polymorphism on the operation of working memory networks has been gained through analysis of brain-based intermediate phenotypes, especially brain activity assessed by neuroimaging. An association of the *COMT* Met allele with better working memory performance has been reported (Barnett et al., 2009; Diaz-Asper et al., 2008; Egan et al., 2001; Goldberg et al., 2003), but the behavioral data have been inconsistent (Barnett, Scorielis, & Munafo, 2008; Bertolino, Blasi et al., 2006; Bruder et al., 2005; Ho et al., 2006; Tsai et al., 2003). Relatively more consistent has been the brain-based finding that the Met allele is associated with indices of greater neural efficiency during working memory tasks (i.e., less recruitment of working memory-associated brain areas for equal behavioral performance; Egan et al., 2001; Mattay et al., 2003; Tan et al., 2007). The inconsistency of the behavioral data highlights the potential utility of the brain-based intermediate phenotype approach for investigating effects of genetic variability that may be obscured at the behavioral level by compensatory factors (e.g., motivation) but visible at the brain level, which is putatively more causally proximate to variations in the genome. The history of psychiatric genetics, which has indicated poor predictive value of initial findings and the tendency for effect sizes to diminish as more studies are published (Flint & Munafo, 2007), suggests continued caution. Nonetheless, the effect of the *COMT* Val158Met polymorphism on brain-based working memory intermediate phenotypes has emerged as one of the best replicated findings in the cognitive neurogenetic literature (Apud et al., 2007; Bertolino, Blasi et al., 2006; Bilder, Volavka, Lachman, & Grace, 2004; de Frias et al., 2009; Egan et al., 2001; Ho et al., 2006; Malhotra et al., 2002; Mattay et al., 2003; Meyer-Lindenberg et al., 2005; Meyer-Lindenberg & Weinberger, 2006; Sambataro et al., 2009; Savitz, Solms, &

Ramesar, 2006; Tan et al., 2007; Winterer & Weinberger, 2004).

One plausible hypothesis that has been proposed to explain the mechanism by which *COMT* and other dopamine-system-related genes influence the function of brain networks involved in working memory centers on the signal-to-noise ratio of dopaminergic signaling within PFC (Apud et al., 2007; Bertolino et al., 2006; Savitz et al., 2006; Winterer et al., 2006; Winterer & Weinberger, 2004). This hypothesis posits that binding at dopamine D1 and D2 receptors affects the signal-to-noise ratio in PFC by modulating the excitatory release of glutamate from pyramidal cells (likely through modulation of Na⁺ channel opening; Rotaru, Lewis, & Gonzalez-Burgos, 2007), and possibly by modulating the inhibitory release of GABA from interneurons (Winterer et al., 2006). In short, the higher the D1/D2 activation ratio, the stronger the excitatory signal, and the better the inhibition of noise in the surround. This is important for working memory, which requires the maintenance of an informational signal over time and inhibition of distractor noise. At levels of dopamine availability that facilitate optimal working memory function, D1 binding in PFC is high relative to D2 binding. However, when dopamine availability is decreased, as it is decreased in Val carriers relative to Met carriers, the D1/D2 ratio is reduced. Thus, the *COMT* Val allele plausibly contributes to a diminished signal-to-noise ratio in PFC and consequently less efficient working memory function.

A related hypothesis, the tonic–phasic hypothesis (de Frias et al., 2009), posits that the low enzyme activity Met allele leads to enhanced tonic dopamine activity in PFC, promoting sustained cognitive representations in working memory, whereas the Val allele leads to reduced tonic but enhanced phasic dopaminergic activity in subcortical regions, putatively supporting greater cognitive flexibility. These hypothesized mechanistic accounts are supported by the finding that the *COMT* Val allele is associated with less efficient activation and reduced functional connectivity in a network of brain regions associated with working memory function, especially DLPFC (Apud et al., 2007; Bertolino, Blasi et al., 2006; Bilder et al., 2004; de Frias et al., 2009; Egan et al., 2001; Ho et al., 2006; Malhotra et al., 2002; Mattay et al., 2003; Meyer-Lindenberg et al., 2005; Meyer-Lindenberg & Weinberger, 2006; Sambataro et al., 2009; Savitz et al., 2006; Tan et al., 2007; Winterer & Weinberger, 2004).

Analysis of dopamine-related performance, however, shows that more dopamine does not always mean better working memory performance. Rather, research in animals and humans has demonstrated that an inverted U-shaped function better describes the relationship between working memory performance and dopamine availability (Cools & Robbins, 2004). Cognitive neurogenetic data concerning the effects of *COMT* on working memory are consistent with this inverted U-shaped function and have added some new insights. Treatments that increase dopamine availability in PFC reduce neural efficiency during working memory in Met allele homozygotes, who have relatively high baseline dopamine availability. Conversely, these treatments improve working memory efficiency in Val homozygotes, whose baseline dopamine availability is lower (Apud et al., 2007; Mattay et al., 2003; Meyer-Lindenberg et al., 2005).

Investigations of *COMT* have provided promising insights regarding the neural mechanisms that may influence working memory. However, it is essential to resist envisioning the association between gene and phenotype (or gene and intermediate phenotype) as a one-to-one relationship (because of pleiotropy and polygenicity; Fisher, 2006; Flint & Munafò, 2007). More comprehensive methodological approaches can be taken to engage the multiplicity of the genotype-intermediate phenotype relationship by examining multiple polymorphisms taken together as predictors of specific neural characteristics (Bertolino, Blasi et al., 2006; de Quervain & Papassotiropoulos, 2006; Meyer-Lindenberg, Nichols et al., 2006; see Fig. 7.2).

For example, Meyer-Lindenberg and colleagues (Meyer-Lindenberg, Nichols et al., 2006) employed a step-wise analysis to examine the combined effect of SNPs at three different loci in the *COMT* gene on neural activity during an N-back working memory task. Consistent with previous findings, Val allele load (0, 1, or 2 Val alleles) at the Val158Met SNP was predictive of prefrontal activation as measured by functional magnetic resonance imaging (fMRI). However, exploratory multipolymorphic analysis indicated that the strongest predictor of prefrontal activation was actually a combination of variants that included all three loci within the *COMT* gene. This analysis also revealed a complex, nonadditive interaction between the Val158Met SNP and one of the other *COMT* SNPs (a P2 promoter region SNP).

These findings point to the complexity of the effects of genotype on brain activity and behavior.

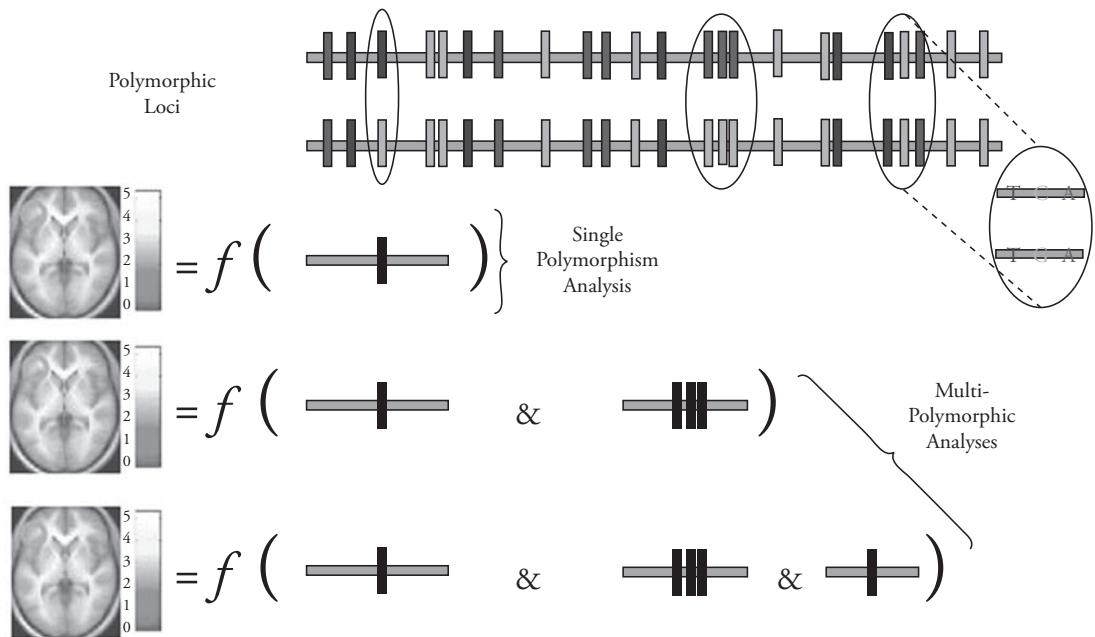


Fig. 7.2 Multipolymorphic analysis. Multiple polymorphic sites in the genome influence a psychological phenotype and its underlying brain activity (brain-based intermediate phenotype). The combined, complexly interactive contributions of these polymorphisms influence neural activity and cognitive function. Most cognitive neurogenetic research to date has focused on variation at individual polymorphisms. These findings are instructive, particularly where only one candidate polymorphism has been implicated by prior findings. However, a growing number of cognitive neurogenetic studies are testing the association of variations at multiple polymorphisms with brain-based intermediate phenotype(s). The figure illustrates a generic multipolymorphic analysis. Polymorphic loci are indicated within two example DNA sequences (where the upper sequence differs from the lower sequence). Brain activity in a region of interest is predicted as a function of one, two, or three of these polymorphisms. Development of multipolymorphic analyses will help construct a better understanding of the complex genetic “networks” that contribute to cognitive brain function and human psychology. See color figure.

Although they generally require larger sample sizes, approaches that consider multiple genetic variations can capture larger fractions of phenotype and intermediate phenotype variance (Nackley et al., 2006) and will be essential to the development of behavioral genetics and cognitive neurogenetics.

Variation in other dopamine-system-related genes has also been linked to working memory performance and related brain function. For example, dopamine transporter (*DAT1*) genotype appears to be associated with response accuracy and activity in frontal, striatal, and parietal regions during working memory tasks (Stollstorff et al., 2010). A separate study indicated that *DAT1* variation influences DLPFC and cingulate activity (Bertolino et al., 2006) during working memory, and that these *DAT1* effects interact with the effects of *COMT* genotype (Bertolino, Blasi, et al., 2006). In a strong example of vertically integrating genetic, molecular-biological, and neural data, Bertolino and colleagues

(Bertolino et al., 2009) have linked variation in the dopamine receptor D2 (*DRD2*) gene to the expression of a particular type of postsynaptic D2 receptor, ligand binding at this receptor, and functional connectivity between striatum and PFC during working memory tasks.

Variations in a number of non-dopaminergic genes also influence the function of working memory networks, including the e4 allele of the apolipoprotein E (*APOE*) gene, which is a risk factor for Alzheimer’s disease (Filbey, Slack, Sunderland, & Cohen, 2006; Wishart et al., 2006), and variations in the glutamate receptor metabotropic 3 (*GRM3*) gene (Egan et al., 2004; Marenco et al., 2006). *GRM3* genotype is associated with DLPFC efficiency and its functional connectivity with parietal cortex (Tan et al., 2007). This finding was revealed by a region-of-interest-based fMRI approach, a useful strategy for minimizing multiple comparisons of brain regions and thus reducing Type I error. Moreover

GRM3 genotype was found to interact with *COMT* genotype, such that homozygosity for the *COMT* Met allele ameliorated the effects of the less efficient *GRM3* genotype (Tan et al., 2007). Variations in the G-protein signaling 4 (*RGS4*) gene, which influences G-protein-linked signal transduction in multiple neurotransmitter systems, have also been implicated in working memory performance as well as functional connectivity and local gray and white matter volume in a network of working memory-related brain regions (Buckholtz et al., 2007).

Polymorphisms in additional brain-expressed genes have been associated with variations in parietal and frontal activity, hippocampal structure and activity, latency of ERP-measured P300 response, and with functional coupling between DLPFC and hippocampus during working memory tasks (Bath & Lee, 2006; Callicott et al., 2005; Cerasa et al., 2010; Egan et al., 2003; Jansen et al., 2009; Reuter et al., 2008; Schofield et al., 2009; Wolf, Jackson, Kissling, Thome, & Linden, 2009). A growing literature surrounding the *BDNF* gene, thus far primarily studied in relation to episodic memory, may prove especially fruitful for insights into working memory and its neural underpinnings (Cerasa et al., 2010; Nagel et al., 2008; Schofield et al., 2009). New behavioral-genetic and cognitive neurogenetic findings related to working memory seem likely to continue emerging rapidly. However, replication of extant findings rather than expansion, will be most effective in establishing a solid understanding of the molecular-genetic mechanisms that support working memory function in PFC.

Attention

Attention is related to working memory. While lower level attention processes (e.g., orienting attention to a surprising noise) can be clearly distinguished from working memory, attention at the executive level (e.g., deciding which of several noises is most important to attend) often is not easily distinguishable from working memory. Tasks that measure working memory (e.g., the N-back task) rely on executive attention, and tasks that measure executive attention, including those discussed later, rely on working memory. Executive control of attention is a major component of working memory theory (Baddeley, 2003; Engle, 2002), and there is overlap between the instantiations of these constructs at the neural level (Kane & Engle, 2002). As such, it is reasonable to predict that a largely shared set of genetic factors impacts these two cognitive phenotypes.

Like working memory, the brain systems related to attention are among the most well characterized and most reliably identified in the cognitive neuroscience literature (Raz & Buhle, 2006). Brain imaging has related attention to large-scale neural networks that are active during attentional tasks and are anatomically overlapping with regions that, when damaged, produce attentional deficits (Mesulam, 1981; Posner & Petersen, 1990). Posner and Petersen proposed that the neural architecture of attention comprises a system of three anatomical networks, which mediate initial alerting of the attentional system, orienting to stimuli, and executive attention, respectively.

Cognitive neurogenetic approaches have been used to test and inform the model of separate attentional networks by examining whether individual differences in distinct components of attention are associated with distinct genetic polymorphisms. Dopaminergic genes have been a focus of this research because of the high levels of expression of dopamine receptors in the anterior cingulate cortex (ACC) and other regions that have been implicated in attention.

The Attention Network Test (ANT) provides a well-validated behavioral measure for parsing the components of attention (Fan, McCandliss, Sommer, Raz, & Posner, 2002) and has been important in establishing the cognitive neurogenetics of attention (see Fig. 7.3). The executive attention measure of the ANT shows about 77% test-retest reliability and a heritability of 89% (Fan, Wu, Fossella, & Posner, 2001). A behavioral-genetic study showed modest associations of executive control of attention with common variants in genes, including the monoamine oxidase-A (*MAO-A*) and the dopamine D4 receptor (*DRD4*) genes (Fossella & Posner, 2004). Neuroanatomical characterizations of brain activity during performance of the ANT (Fan & Posner, 2004) revealed dissociable networks for the three putative components of attention as well as evidence that individual differences in the activation of cingulate cortex correlates with variation in the *MAO-A* gene (Fan, Fossella, Sommer, Wu, & Posner, 2003). This *MAO-A* finding has subsequently been independently replicated and extended (Meyer-Lindenberg et al., 2006).

Because the executive component of attention is an important aspect of working memory (Baddeley, 2003; Engle, 2002), it is not surprising that common genetic variants are associated with both working memory and executive attention.

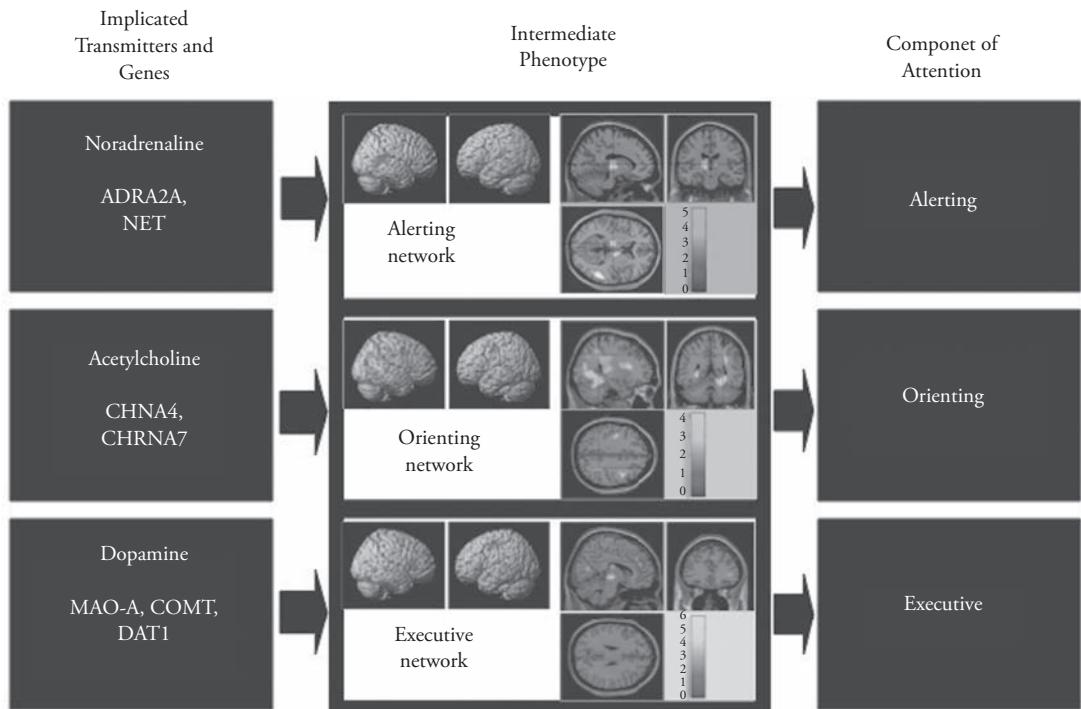


Fig. 7.3 Parsing components of attention in cognitive neurogenetic studies with the Attention Networks Test (ANT). Variations in a set of candidate genes have been correlated with behavioral performance or brain activity for distinct components of attention using the ANT. To generate candidate genes for cognitive neurogenetic studies—in this case for studies on attention—it is important to rely on multiple sources of converging evidence. First, neuroimaging and lesion data pointed to separable neural networks that carry out different aspects of attention. Secondly, pharmacological manipulations demonstrated that noradrenergic modulation can influence the efficiency of the alerting network, while cholinergic modulation and dopaminergic modulation can influence orienting and executive control of attention, respectively. With regard to neural intermediate phenotypes, brain networks were identified where attention-related activity overlaps with patterns of gene expression. It was hypothesized that variation in gene sequence should correlate with individual differences in neural activity associated with components of attention. In the case of executive control of attention, hypotheses of this kind have been supported for brain regions that are targets of dopamine innervation such as the frontal midline, lateral prefrontal areas, and basal ganglia. See color figure.

For example, during tasks that demand executive attention, *COMT* Met allele carriers show relatively decreased activity in ACC and perform better (Blasi et al., 2005; Krabbendam et al., 2006; Winterer et al., 2006), suggesting that the Met allele is associated with more efficient neural processing during such tasks. However, as with working memory, the data relating polymorphic variation to behavior have been less consistent than the data linking polymorphic variation to individual differences in neural intermediate phenotype (Reuter, Ott, Vaitl, & Hennig, 2007), again indicating the value of the brain-based intermediate phenotype approach. For example, Blasi et al. (2005) demonstrated that the effect of *COMT* genotype on ACC activation is strongest at the highest levels of executive attentional demand. The effects of *COMT* genotype on executive attention

have been framed under the dopamine signal-to-noise hypothesis (Winterer et al., 2006; Winterer & Weinberger, 2004), and increased levels of noise have been demonstrated in Val allele carriers surrounding peaks of activation in ACC and elsewhere in PFC (Winterer et al., 2006). Extraction of *COMT* protein from cadaverous human tissue (Tunbridge et al., 2007) and a recent study of *COMT* attentional effects in toddlers (Voelker, Sheese, Rothbart, & Posner, 2009) suggest that the *COMT* enzymatic activity and the posited cognitive effects of *COMT* genotype may increase substantially from early childhood to adulthood.

Variants in dopamine-system-related genes other than *COMT* also affect the operation of attentional brain networks as a function of candidate genetic variants. Early, and still somewhat contradictory findings have begun to implicate several other

dopamine-system-related polymorphisms in the executive component of attention based on performance measures, evoked potentials, and functional imaging of ACC activation (Birkas et al., 2006; Fan et al., 2003; Fossella, Green, & Fan, 2006; Fossella et al., 2002; Passamonti et al., 2008; Passamonti et al., 2006; Rueda, Rothbart, McCandliss, Saccomanno, & Posner, 2005).

While dopamine-system-related genes have been associated with the executive component of attention, a number of genes that are not directly involved in the dopamine system have been associated with the orienting and alerting components of attention and their associated brain networks. These include genes associated with the glutamatergic (Gallinat et al., 2007), GABA-ergic (Porjesz et al., 2002), serotonergic (Fallgatter et al., 2004), and cholinergic (Begleiter & Porjesz, 2006; Espeseth et al., 2006; Parasuraman, Greenwood, Kumar, & Fossella, 2005; Winterer et al., 2007) systems. Cholinergic-system-related genes are neurophysiologically plausible candidates because of the role of acetylcholine in activating parietal attention networks in response to salient stimuli (Davidson & Marrocco, 2000; Sarter & Bruno, 1997). Indeed, both nicotinic and muscarinic acetylcholine receptor genes have been associated with attentional alerting as indexed by individual differences in response time, as well as brain-based intermediate phenotypic measures obtained by electrical recording, and brain imaging of alerting tasks (Begleiter & Porjesz, 2006; Espeseth et al., 2006; Greenwood, Lin, Sundararajan, Fryxell, & Parasuraman, 2009; Parasuraman et al., 2005; Winterer et al., 2007). These findings, juxtaposed with those implicating dopamine-system-related genes in the executive component of attention (and working memory), reinforce the separable networks model of attention (Posner & Petersen, 1990; Posner, Rothbart, & Sheese, 2007; Raz & Buhle, 2006) and enrich this model by identifying separable genetic influences that bear upon distinct brain networks.

Long-Term Memory

Long-term, declarative memory is a major focal area for neuroscience research. Whereas attention and working memory generally operate over brief durations, long-term memory requires neural mechanisms and networks that can represent information over extended durations. Thus, mechanistic models of memory that incorporate long-term changes in synaptic strength have gained empirical

and theoretical traction. Synaptic plasticity as a mechanism of memory has been supported by findings at the cellular and systems level (C. Chen & Tonegawa, 1997; Silva, 2003). The molecular pathways mediating synaptic plasticity and the neural architecture mediating memory function have been characterized in some detail (Chen & Tonegawa, 1997; Lamprecht & LeDoux, 2004). Furthermore, genetic investigation of memory performance has revealed roughly 50% trait heritability (McClearn et al., 1997). Thus, research on long-term memory is well suited for cognitive neurogenetic approaches.

A growing body of work evaluates the role of the *BDNF* gene, which encodes the BDNF protein. This protein shows high levels of brain expression in hippocampal regions (Murer, Yan, & Raisman-Vozari, 2001), which are critical for long-term memory function (Gabrieli, 1998). Evidence from animal models indicates that BDNF-mediated long-term potentiation (LTP) in the hippocampus facilitates memory formation (Poo, 2001). In human cells, the Met allele of a common Val-to-Met substitution polymorphism at codon 66 of the *BDNF* polypeptide is associated with diminished BDNF secretion (Chen et al., 2004; Egan et al., 2003). This effect is likely due to disrupted trafficking of the BDNF protein through the secretory pathway within the cell (Chen et al., 2004; Egan et al., 2003), which may result from decreased interaction between BDNF and the cytoplasmic trafficking protein, Sortillin (Chen et al., 2005).

Because *BDNF* genotype has been implicated as an influence on synaptic plasticity in hippocampus (Goodman et al., 1996; Kang & Schuman, 1995; Poo, 2001), a number of studies have begun to ask whether the role of *BDNF* can be related to models of memory in which cellular plasticity is a core mechanism. Structural brain imaging has shown that the Met allele is reliably associated with reduced hippocampal and parahippocampal volume in healthy adults (Bueller et al., 2006; Ho et al., 2006; Pezawas et al., 2004; Szeszko et al., 2005), although these effects may be due to global effects of the *BDNF* Val66Met polymorphism on brain volume (Toro et al., 2009). *BDNF* Met allele carriers show poorer cognitive performance on tasks such as episodic recall of scenes and events (Dempster et al., 2005; Egan et al., 2003; Hariri et al., 2003) that typically rely on hippocampal activation. Some evidence indicates that verbal memory tasks, such as remembering lists of words, do not show strong effects of *BDNF* genotype (Bath & Lee, 2006), possibly because they are less reliant on hippocampal mechanisms (Bath & Lee, 2006;

Egan et al., 2003). However, at least one study has identified impaired performance among healthy Met allele carriers across a range of verbal memory tasks (Ho et al., 2006). The Met allele has also been associated with decreased hippocampal engagement at both the encoding and retrieval stages of episodic memory (Bath & Lee, 2006; Callicott et al., 2005; Hariri et al., 2003; Hashimoto et al., 2008), and decreased hippocampal levels of n-acetyl-aspartate, a measure of in vivo synaptic activity (Callicott et al., 2005; Egan et al., 2003).

Although the cellular pathways related to *BDNF* and synaptic plasticity are attractive candidates for translational research on cognitive mechanisms, memory is a complex, multiphasic process. As such, it is affected by a variety of genetic factors that exert mechanistic effects on a range of processes, including not only storage but also encoding and retrieval. These genetic effects are observable in the operation of a network of brain regions that is centered on the hippocampus but depends on connectivity with frontal and midbrain regions. An expanding literature has begun to characterize the effects of a number of specific candidate genes on several components of the networks that support memory. These include genes whose protein products, like *BDNF*, are linked to synaptic plasticity, such as the *CPEB* gene encoding the cytoplasmic polyadenylation element-binding protein (Vogler et al., 2009) and the *PRNP* gene encoding prion (Papassotiropoulos, Wollmer et al., 2005). Individual differences in memory performance have also been linked to variation in the *DISC1* gene, a schizophrenia susceptibility factor (Callicott et al., 2005), and schizophrenia-linked alleles at three polymorphisms of the neurotrophin receptor (*NTRK-3*) gene have been associated with differences in hippocampal activity during episodic encoding (Ottness et al., 2009).

As in the working memory and attention literatures, genes that influence dopaminergic signaling have been implicated. Met genotype at the *COMT* Val158Met SNP has been associated with stronger hippocampal activation during encoding and retrieval of remembered information (Bertolino, Rubino, et al., 2006; Krach et al., 2010) and with better declarative memory performance in adults and children (Bertolino, Rubino et al., 2006; de Frias et al., 2004; Diamond, Briand, Fossella, & Gehlbach, 2004). *COMT* Met allele load also correlates with a putatively advantageous decrease in the functional coupling of hippocampal and

ventral-lateral prefrontal cortex activity (Bertolino, Rubino et al., 2006). Dorsal and orbital prefrontal regions show stronger functional coupling with hippocampus among Met homozygotes, which was linked to successful encoding in a verbal memory paradigm (Schott et al., 2006). This *COMT*-related prefrontal intermediate phenotype was differentiated from an effect of *DAT1* genotype on midbrain and ACC activation during the same verbal memory task, possibly indicating a role for *DAT1* in the rapid, phasic midbrain activity that is associated with memory demands (Schott et al., 2006).

Neurotransmitter-related polymorphisms not directly related to dopamine have also been associated with memory performance and the function of memory-relevant neural systems. These include variations in the serotonin-2A receptor (*5-HT2A*) gene (de Quervain et al., 2003; Papassotiropoulos et al., 2005) and the alpha2b-adrenergic receptor (*ADRA2B*) gene, which has been specifically linked to emotional memories and amygdala activation during encoding (de Quervain et al., 2007; Rasch et al., 2009).

Several insights into genetic factors that affect the molecular- and systems-level implementation of memory have arisen from studies of a panel of genes that influence memory-related cellular signaling and synaptic plasticity. de Quervain and Papassotiropoulos (2006) targeted a specific cluster of genes expressed within a cellular signaling cascade that includes the metabotropic glutamate receptor and the adenylyl cyclase enzyme. A large group of behaviorally tested participants was assigned genotype "scores" based on their genotype at seven polymorphisms in this genetic cluster. Thirty-two participants whose behavioral performance was similar were selected for an fMRI study. This approach enabled the researchers to isolate the effects of genotype on intermediate phenotype while controlling for other factors related to behavioral performance that might influence brain activity (e.g., motivation). Genotype scores that correlated with greater hippocampal and medial temporal activity in fMRI participants were also associated with better memory performance among non-fMRI participants.

These researchers similarly demonstrated the effect of a polymorphism in the calmodulin-binding transcription activator 1 (*CAMTA1*) gene on medial temporal and hippocampal activity during a task of episodic memory retrieval (Huentelman et al., 2007). This *CAMTA1* polymorphism was first identified as

a candidate by a high-density genome-wide scan of 502,627 SNPs on a DNA microarray (frequently referred to as a “gene chip”) to screen for SNPs that were predictive of memory performance among a large sample of behaviorally tested participants (Huentelman et al., 2007). A tyrosine-to-cytosine substitution in the *KIBRA* gene was also a predictor of better episodic association memory in this population. At the neural level, *KIBRA* genotype influences the efficiency of memory-related hippocampal activation (Papassotiropoulos et al., 2006) as well as hippocampal volume and functional coupling between hippocampal and cingulate cortices (Emery et al., 2009). These effects are putatively due to an effect of KIBRA, a novel cytoplasmic protein, on synaptic plasticity that supports long-term potentiation in hippocampus (Emery et al., 2009; Papassotiropoulos et al., 2006). DNA microarray analysis of the kind utilized in this and related research (Potkin et al., 2009) is a powerful tool for identifying candidate polymorphisms in broad swaths of the genome, and it is likely to have great impact on future cognitive neurogenetic investigation. In addition, the strategies employed by Papassotiropoulos et al. further demonstrate the usefulness of considering multiple polymorphisms in the investigation of single cognitive phenotypes in order to begin engaging the extensive polygenicity of human cognitive traits (de Quervain & Papassotiropoulos, 2006; Rasch, Papassotiropoulos, & de Quervain, 2010).

Language

Understanding the underpinnings of language function has been a topic of great interest in cognitive neuroscience, and it has great importance for informing the treatment of language disorders. However, language processing has received relatively little attention in the cognitive neurogenetics literature, despite a relatively well-delineated neural architecture (Binder et al., 1997; Hickok & Poeppel, 2007). The genetics of language function have been approached through linkage studies, which have yielded important results concerning the chromosomal loci of language-related genes (Bishop, 2001; Fisher & DeFries, 2002; Fisher, Lai, & Monaco, 2003). Structural imaging has demonstrated high heritability of gray matter volume in prefrontal language areas as compared to other prefrontal regions (Thompson et al., 2001). However, only a few of cognitive neurogenetic investigations have sought to tie language-related brain networks to specific genes implicated in language function. In

particular, more work is needed to identify molecular-genetic and systems-level mechanisms that contribute to variation in normal, nondisordered language processing.

Much of the research performed to date has focused on a point mutation in the forkhead box P2 (*FOXP2*) gene, a transcription factor that regulates genes involved in neural development. *FOXP2* contains evolutionarily recent amino-acid changes that may have arisen with the advent of human spoken language (Enard et al., 2002; Fisher et al., 2003; Lai, Gerrelli, Monaco, Fisher, & Copp, 2003). Variation in the *FOXP2* gene has been linked to a rare language disorder in family studies among patients and their unaffected relatives (Fisher et al., 2003; Lai, Fisher, Hurst, Vargha-Khadem, & Monaco, 2001). Variation in *FOXP2* genotype has been associated with gray matter density in several areas involved in language production as well as language reception and comprehension (Belton, Salmond, Watkins, Vargha-Khadem, & Gadian, 2003; Watkins et al., 2002). People with the nondisordered (normal) *FOXP2* genotype showed stronger recruitment than affected siblings in Broca’s area and inferior prefrontal regions associated with language comprehension during both overt and covert verb generation (Liegeois et al., 2003).

A novel candidate polymorphism, rs1006737 in the alpha 1C subunit of the L-type voltage-gated calcium channel (*CACNA1C*) gene, which has been previously implicated in impaired verbal fluency in bipolar disorder, has recently been shown to affect inferior frontal gyrus activity during a semantic fluency task in healthy participants and to affect performance of a more demanding semantic fluency task performed outside the scanner (Krug et al., 2010). The same group of researchers has demonstrated that inferior frontal and right middle temporal activation, and out-of-scanner semantic fluency performance, decrease in healthy participants concomitant with increasing numbers of the schizophrenia-risk-associated T allele of the rs35753505 SNP of the neuregulin 1 (*NRG1*) gene (Kircher et al., 2009). These researchers have also shown an association between genotype at rs1018381 of the schizophrenia-linked dysbindin 1 (*DTNBP1*) gene and neural efficiency in ACC and superior and middle temporal gyrus (Markov et al., 2009), as well as an association between Val allele load at the *COMT* Val158Met SNP and left inferior frontal activity (Krug et al., 2009), during their semantic fluency task.

Additional insight into the cognitive neurogenetics of language has come from an investigation of the *PLXNB3* gene, which is expressed predominantly in the brain and encodes Plexin B3, a potent stimulator of neurite growth (Rujescu et al., 2007). Two candidate polymorphisms on the *PLXNB3* gene were used to define three distinct haplotypes. Carriers of the most evolutionarily recent haplotype scored highest for verbal intelligence and showed the greatest whole-brain white matter volume (Rujescu et al., 2007). Genetic influences on neurite growth and white matter connectivity have also been implicated in normal language-related variation in studies examining the effects of *BDNF* genotype on verbal cognition (Egan et al., 2003; Ho et al., 2006). In addition, the *CACNA1C* gene has been implicated in mediation of synaptic plasticity via the encoded calcium receptor's role in a second messenger cascade (Moosmang et al., 2005; White et al., 2008), and the *NRG1* gene is linked to white matter myelination (Nave & Salzer, 2006). These initial findings provide a starting point for future investigations to directly evaluate whether molecular-genetic determinants of white matter connectivity and efficacy account for variance in functional brain-based measures of language.

Conclusions and Future Directions

Connecting variations in the genome to individual differences in human brain function is not only an endeavor for clinicians. This approach has begun to inform cognitive neuroscientific models of variation among healthy individuals. It is early yet, but coalescing themes can be discerned in data. Mental abilities such as working memory and attention that are characterized by fast, nimble responses to dynamic stimuli, seem to be most influenced by genetic variables related to neurotransmitter signaling (e.g., dopamine transmission), a relatively rapid-fire form of communication in the brain. By contrast, long-term memory seems to show a relatively strong influence of genes (e.g., *BDNF*) that shape long-duration synaptic connections between neurons. Meanwhile, early returns from the cognitive neurogenetics of language implicate genes (e.g., *FOXP2*) that guide construction of the developing brain. These broad themes may become more definite under closer scrutiny, and with further aggregation of data, or they may dissolve into finer-grained hypotheses delineating individual causal pathways from genes to brain to cognition.

New Data, Familiar Theory

Molecular-genetic data can contribute to cognitive neuroscience in multiple ways. They are a source of key constraints on the mechanisms that mediate a given psychological function, providing a qualitatively different measure with which to test hypotheses. This function is similar to the way in which neural data have helped inform cognitive science by providing empirical constraints, which help to test competing hypotheses about the implementation of a cognitive function. Molecular-genetic data can also be used to identify causes of individual variation (specific genes and gene–environment interactions), which can both help to explain such variation, and to remove unwanted variation from cognitive and brain-based investigations that take genotype as a covariate (Green et al., 2008). Genetic variation can be leveraged to parse and inform between-group differences in neural structure and function that correlate with genetic variables of interest. This use of genetic research is similar to the way that cognitive neuroscience data help to parse and inform individual differences in psychological constructs (e.g., intelligence, personality, clinical status). Brain-based indices are now taken to be physical markers of individual differences in cognitive function and to contain new information about mechanism not visible in behavior alone, sometimes even in the absence of behavioral differences (Gray et al., 2005). The potential exists for cognitive neurogenetic data to identify even deeper-lying markers and to establish the scaffolding for a vertical integration of all data concerning a cognitive function, extending from psychological phenotype down through biomolecular implementation and genomic coding.

Such potential can be glimpsed in the literature that is developing around a number of the genetic variants reviewed here. In these cases, cognitive neuroscience in humans has pointed toward neural systems and characteristics that support a particular cognitive function (e.g., hippocampal plasticity in declarative memory), and candidate genetic variants (e.g., the *BDNF* Val66Met polymorphism) have been selected that are plausibly related to the identified neural systems. As neural and behavioral effects (e.g., differential hippocampal activation and memory performance) come to reliably implicate molecular-genetic mechanisms “downstream” of the targeted genetic variants (e.g., putatively differential trafficking and secretion of *BDNF*), this information will be used to inform and constrain subsequent models and predictions.

More immediately, genetic data provide an additional level of analysis to complement and inform traditional cognitive neuroscience questions. For example, in the attention literature, the use of genotypic dissociations to enrich behavioral and neural dissociations between types of attention is a relatively well-developed demonstration of the utility of genetic data in cognitive neuroscience.

Indications for Future Research

Much as cognitive neuroscientific data have enriched earlier behavioral models of the mechanisms supporting psychological function, one might expect a similar enrichment of cognitive neuroscience over the next 20 years with the incorporation of genetic data. However, an understanding based on genetic dissociations, or gene–brain associations, without careful regard for psychological theory, will not be enough. The ubiquity and strength of gene-environment interactions requires that a complete description also account for the environment. The environment includes any contextual effects on gene function whether these be other genes, developmental influences, or the wider social-psychological milieu (Caspi & Moffitt, 2006; Fisher, 2006). In addition, as new candidate polymorphisms continue to be identified, meaningfully integrating this information with the cognitive neuroscience literature will require more than fragmentary associations with individual cognitive tasks. Grouping these associations in principled ways, and testing genetic effects on underlying cognitive constructs—constrained by predictions based on brain regions known to underlie those constructs—will be essential to finding constellations of meaning on a horizon already crowding with new and sometimes seemingly contradictory genetic associations. Thus, psychological theory and psychometric expertise will be more indispensable, not less, for channeling an influx of a new type of data toward the familiar goal of understanding how properties of mind are constituted physically (Plomin, 2000).

NOT JUST IMAGING

Full development of a young field will depend on the use of multiple cognitive neuroscience methods—not just brain imaging—in order to overcome the inherent limitations of any single method. For example, measuring the effects of cholinergic-system-related gene variation on electrical oscillation patterns in the brain (e.g., Begleiter & Porjesz, 2006, or the effects of serotonin-system-related genetic

variation on neural error processing at the millisecond scale (e.g., Fallgatter et al., 2004), requires direct sensitivity to electrical activity, which fMRI and other brain-imaging methods cannot provide. In addition, agonist and other chemical treatment studies in clinical and healthy populations, cadaverous tissue expression assays, temporarily induced lesion studies, and other brain-based methods will help to provide breadth and convergence of data to grow the field and strengthen its core findings. For this reason we prefer the more comprehensive term, *cognitive neurogenetics*, to the more focused term, *imaging genetics*.

RULES OF THUMB

At least four specific prescriptive indications for future investigations can be gleaned from the early history of cognitive neurogenetics.

1) Use large sample sizes, much larger than typical sample sizes for MRI, PET, and EEG. Much, if not most, cognitive neurogenetic work to date has been underpowered to accurately reject the null hypothesis (Green et al., 2008). De Geus and colleagues propose a rule of thumb of at least 10 participants in the smallest genotype group (de Geus, Goldberg, Boomsma, & Posthuma, 2008), which will often require sample sizes in the hundreds. Higher numbers of participants are also required in order to employ DNA microarray analyses in studies of brain-assayed participants.

2) Adhere to psychological theory and psychometric rigor. Points of emphasis here include the use of cognitive tasks with strong construct validity, the use of latent variables comprising multiple theoretically related tasks to represent a psychological construct, and replication of effects on closely related tasks within and between laboratories.

3) Seek integration with molecular-genetic and molecular-biological data. In particular, what is known about regional brain expression and protein-level effects (e.g., enzymatic activity) of a genetic variation should constrain a priori hypotheses about brain characteristics targeted for investigation as intermediate phenotypes. As much as possible, all the levels of measurement in a particular study (e.g., genotype, transmission/binding, neural activity, brain structure, cognitive performance) should be integrated into a single model to test directional predictions about the effects of genetic variation on downstream variables and the effects of downstream variables on each other.

4) Take a systems approach that begins to engage the complexity of genetic networks (variations in different genes and multiple polymorphisms in the same gene) and their effects on networks of multiple, networking brain regions.

A PATH-ANALYTIC APPROACH

A statistical framework that begins to address these indications (although imperfectly) is path-analytic mediation modeling. Path-analytic mediation models have traditionally been applied to investigate relationships between genetic and environmental variables (e.g., Munafo, 2006). Somewhat surprisingly, however, they have not yet been widely used to test whether relationships between genes and cognition are mediated by intermediate neural phenotypes. Cognitive neurogenetic studies have typically examined pairwise relationships (e.g., gene-brain and gene-cognition correlations), but pairwise relationships cannot test statistical mediation. Thus, they cannot test whether variation in a neural intermediate phenotype variable mediates the effect of a genetic variable on a cognitive variable. We have recently constructed a path-analytic mediation model that is directional from gene to brain to cognition, using data collected from 160 genotyped fMRI subjects who performed the multisource interference (MSIT) task of executive attention (Green, Kraemer, DeYoung, Fossella, & Gray, unpublished data). This model indicates an effect of *COMT* genotype on MSIT performance that is mediated by activity in a group of frontal brain regions.

Beyond testing neural mediation hypotheses, path-analytic mediation models in cognitive neurogenetics have several attractive attributes. (1) They conceptually articulate the expression pathway (gene-to-brain-to-cognition) in a clear and integrated way that is consistent with directional/causal hypotheses. This is appropriate because, with the possible exception of epigenetic influences, genetic effects generally cause neural and cognitive effects, not vice versa. (2) They can incorporate multiple intermediate phenotype variables simultaneously (e.g., activity in multiple brain regions). This is useful because cognitive functions usually involve multiple co-contributing regions and structures. (3) They can expand to include additional variables and complexity as new data inform existing models (e.g., new evidence concerning epigenetic moderator variables). (4) They organically incorporate latent variables comprising shared variance between multiple measures (e.g., multiple cognitive tasks which, together, represent an underlying psychological construct).

To achieve sufficient statistical power, path-analytic mediation requires sample sizes that may appear dauntingly large by the standards of brain-imaging research, which relies on methods that are expensive and time consuming. This is likely why mediation models have not been widely used in cognitive neurogenetics to date. The requirement of a large sample is a considerable challenge in practice; however, it is not a problem in principle. Moreover, this requirement reflects the broader reality that progress in cognitive neurogenetic research will require large-scale data collections.

In Sum

For cognitive neuroscience, the promise of molecular genetics is one of growth from characterizing the “where” and “when” of mental function toward new questions of “whence” and “how,” using a new set of powerful tools that can reveal mechanistic insights at the level of fundamental biology. While this promise is still largely unproven, and interpretive hazards require great caution, the research reviewed here suggests some emerging proof of concept and grounds for measured optimism.

Acknowledgments

The authors wish to thank M.I. Posner, J. Fossella, J. Gray, M. Munafo, and C. G. DeYoung, whose consultation contributed much to the manuscript.

References

- Allen, E. G., Sherman, S., Abramowitz, A., Leslie, M., Novak, G., Rusin, M., ... Letz, R. (2005). Examination of the effect of the polymorphic CCG repeat in the FMRI gene on cognitive performance. *Behavior Genetics*, 35(4), 435–445.
- Apud, J. A., Mattay, V., Chen, J., Kolachana, B. S., Callicott, J. H., Rasetti, R., ... Weinberger, R. (2007). Tolcapone improves cognition and cortical information processing in normal human subjects. *Neuropsychopharmacology*, 32(5), 1011–1020.
- Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, 4(10), 829–839.
- Barnett, J. H., Heron, J., Goldman, D., Jones, P. B., & Xu, K. (2009). Effects of catechol-O-methyltransferase on normal variation in the cognitive function of children. *American Journal of Psychiatry*, 166(8), 909–916.
- Barnett, J. H., Scorielis, L., & Munafo, M. R. (2008). Meta-analysis of the cognitive effects of the catechol-O-methyltransferase gene val158/108met polymorphism. *Biological Psychiatry*, 64, 137–144.
- Bath, K. G., & Lee, F. S. (2006). Variant BDNF (Val66Met) impact on brain structure and function. *Cognitive, Affective, and Behavioral Neuroscience*, 6(1), 79–85.
- Begleiter, H., & Porjesz, B. (2006). Genetics of human brain oscillations. *International Journal of Psychophysiology*, 60(2), 162–171.
- Belton, E., Salmon, C. H., Watkins, K. E., Vargha-Khadem, F., & Gadian, D. G. (2003). Bilateral brain abnormalities

- associated with dominantly inherited verbal and orofacial dyspraxia. *Human Brain Mapping*, 18(3), 194–200.
- Bertolino, A., Blasi, G., Latorre, V., Rubino, V., Rampino, A., Sinibaldi, L.,...Dallapiccola, B. (2006). Additive effects of genetic variation in dopamine regulating genes on working memory cortical activity in human brain. *Journal of Neuroscience*, 26(15), 3918–3922.
- Bertolino, A., Fazio, L., Di Giorgio, A., Blasi, G., Romano, R., Taurisano, P.,...Sadee, W. (2009). Genetically determined interaction between the dopamine transporter and the D2 receptor on prefronto-striatal activity and volume in humans. *Journal of Neuroscience*, 29(4), 1224–1234.
- Bertolino, A., Rubino, V., Sambataro, F., Blasi, G., Latorre, V., Fazio, L.,...Scarabino, T. (2006). Prefrontal-hippocampal coupling during memory processing is modulated by COMT val158met genotype. *Biological Psychiatry*, 60(11), 1250–1258.
- Bilder, R. M., Volavka, J., Lachman, H. M., & Grace, A. A. (2004). The catechol-O-methyltransferase polymorphism: relations to the tonic-phasic dopamine hypothesis and neuropsychiatric phenotypes. *Neuropsychopharmacology*, 29(11), 1943–1961.
- Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., & Prieto, T. (1997). Human brain language areas identified by functional magnetic resonance imaging. *Journal of Neuroscience*, 17(1), 353–362.
- Birkas, E., Horvath, J., Lakatos, K., Nemoda, Z., Sasvari-Szekely, M., Winkler, I., & Gervai, J. (2006). Association between dopamine D4 receptor (DRD4) gene polymorphisms and novelty-elicited auditory event-related potentials in preschool children. *Brain Research*, 1103(1), 150–158.
- Bishop, D. V. (2001). Genetic influences on language impairment and literacy problems in children: Same or different? *Journal of Child Psychology and Psychiatry*, 42(2), 189–198.
- Bishop, S. J., Fossella, J., Croucher, C. J., & Duncan, J. (2008). COMT val158met genotype affects recruitment of neural mechanisms supporting fluid intelligence. *Cerebral Cortex*, 18(9), 2132–2140.
- Blackwood, D. H., Fordyce, A., Walker, M. T., St Clair, D. M., Porteous, D. J., & Muir, W. J. (2001). Schizophrenia and affective disorders—cosegregation with a translocation at chromosome 1q42 that directly disrupts brain-expressed genes: Clinical and P300 findings in a family. *American Journal of Human Genetics*, 69(2), 428–433.
- Blasi, G., Mattay, V. S., Bertolino, A., Elvevag, B., Callicott, J. H., Das, S.,...Weinberger, D. R. (2005). Effect of catechol-O-methyltransferase val158met genotype on attentional control. *Journal of Neuroscience*, 25(20), 5038–5045.
- Bouchard, T. J., Jr., & McGue, M. (1981). Familial studies of intelligence: A review. *Science*, 212(4498), 1055–1059.
- Bruder, G. E., Keilp, J. G., Xu, H., Shikhman, M., Schori, E., Gorman, J. M., & Gilliam, T. C. (2005). Catechol-O-methyltransferase (COMT) genotypes and working memory: associations with differing cognitive operations. *Biological Psychiatry*, 58(11), 901–907.
- Buckholtz, J. W., Meyer-Lindenberg, A., Honea, R. A., Straub, R. E., Pezawas, L., Egan, M. F.,...Callicott, J. H. (2007). Allelic variation in RGS4 impacts functional and structural connectivity in the human brain. *Journal of Neuroscience*, 27(7), 1584–1593.
- Bueller, J. A., Aftab, M., Sen, S., Gomez-Hassan, D., Burmeister, M., & Zubietta, J. K. (2006). BDNF Val66Met allele is associated with reduced hippocampal volume in healthy subjects. *Biological Psychiatry*, 59(9), 812–815.
- Callicott, J. H., Straub, R. E., Pezawas, L., Egan, M. F., Mattay, V. S., Hariri, A. R.,...Weinberger, D. R. (2005). Variation in DISC1 affects hippocampal structure and function and increases risk for schizophrenia. *Proceedings of the National Academy of Sciences USA*, 102(24), 8627–8632.
- Canli, T., Qiu, M., Omura, K., Congdon, E., Haas, B. W., Amin, Z.,...Lesch, K. P. (2006). Neural correlates of epigenesis. *Proceedings of the National Academy of Sciences USA*, 103(43), 16033–16038.
- Caspi, A., & Moffitt, T. E. (2006). Gene-environment interactions in psychiatry: Joining forces with neuroscience. *Nature Reviews Neuroscience*, 7(7), 583–590.
- Caspi, A., Sugden, K., Moffitt, T. E., Taylor, A., Craig, I. W., Harrington, H.,...Poulton, R. (2003). Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science*, 301(5631), 386–389.
- Cerasa, A., Tongiorgi, E., Fera, F., Gioia, M. C., Valentino, P., Liguori, M.,...Quattrone, A. (2010). The effects of BDNF Val66Met polymorphism on brain function in controls and patients with multiple sclerosis: an imaging genetic study. *Behavioural Brain Research*, 207(2), 377–386.
- Chen, C., & Tonegawa, S. (1997). Molecular-genetic analysis of synaptic plasticity, activity-dependent neural development, learning, and memory in the mammalian brain. *Annual Review of Neuroscience*, 20, 157–184.
- Chen, Z. Y., Ieraci, A., Teng, H., Dall, H., Meng, C. X., Herrera, D. G.,...Lee, F. S. (2005). Sortilin controls intracellular sorting of brain-derived neurotrophic factor to the regulated secretory pathway. *Journal of Neuroscience*, 25(26), 6156–6166.
- Chen, Z. Y., Patel, P. D., Sant, G., Meng, C. X., Teng, K. K., Hempstead, B. L., & Lee, F. S. (2004). Variant brain-derived neurotrophic factor (BDNF) (Met66) alters the intracellular trafficking and activity-dependent secretion of wild-type BDNF in neurosecretory cells and cortical neurons. *Journal of Neuroscience*, 24(18), 4401–4411.
- Cools, R., & Robbins, T. W. (2004). Chemistry of the adaptive mind. *Philosophical Transactions of the Royal Society A: Math, Physics, and Engineering Sciences*, 362(1825), 2871–2888.
- Davidson, M. C., & Marrocco, R. T. (2000). Local infusion of scopolamine into intraparietal cortex slows covert orienting in rhesus monkeys. *Journal of Neurophysiology*, 83(3), 1536–1549.
- de Frias, C. M., Annerbrink, K., Westberg, L., Eriksson, E., Adolfsson, R., & Nilsson, L. G. (2004). COMT gene polymorphism is associated with declarative memory in adulthood and old age. *Behavior Genetics*, 34(5), 533–539.
- de Frias, C. M., Marklund, P., Eriksson, E., Larsson, A., Oman, L., Annerbrink, K.,...Nyberg, L. (2009). Influence of COMT gene polymorphism on fMRI-assessed sustained and transient activity during a working memory task. *Journal of Cognitive Neuroscience*, 22(7), 1614–1622.
- de Geus, E., Goldberg, T., Boomsma, D. I., & Posthuma, D. (2008). Imaging the genetics of brain structure and function. *Biological Psychology*, 79(1), 1–8.
- de Quervain, D. J., Henke, K., Aerni, A., Coluccia, D., Wollmer, M. A., Hock, C.,...Papassotiropoulos, A. (2003). A functional genetic variation of the 5-HT2a receptor affects human memory. *Nature Neuroscience*, 6(11), 1141–1142.
- de Quervain, D. J., Kolassa, I. T., Erd, V., Onyut, P. L., Neuner, F., Elbert, T., & Papassotiropoulos, A. (2007). A deletion variant of the alpha2b-adrenoceptor is related to emotional memory in Europeans and Africans. *Nature Neuroscience*, 10(9), 1137–1139.

- de Quervain, D. J., & Papassotiropoulos, A. (2006). Identification of a genetic cluster influencing memory performance and hippocampal activity in humans. *Proceedings of the National Academy of Sciences USA*, 103(11), 4270–4274.
- Dempster, E., Toulopoulou, T., McDonald, C., Bramon, E., Walshe, M., Filbey, F., et al. (2005). Association between BDNF val66 met genotype and episodic memory. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 134(1), 73–75.
- Diamond, A., Briand, L., Fossella, J., & Gehlbach, L. (2004). Genetic and neurochemical modulation of prefrontal cognitive functions in children. *American Journal of Psychiatry*, 161(1), 125–132.
- Diaz-Asper, C. M., Goldberg, T. E., Kolachana, B. S., Straub, R. E., Egan, M. F., & Weinberger, D. R. (2008). Genetic variation in catechol-O-methyltransferase: effects on working memory in schizophrenic patients, their siblings, and healthy controls. *Biological Psychiatry*, 63(1), 72–79.
- Dick, D. M., Aliev, F., Kramer, J., Wang, J. C., Hinrichs, A., Bertelsen, S., et al. (2007). Association of CHRM2 with IQ: converging evidence for a gene influencing intelligence. *Behavior Genetics*, 37(2), 265–272.
- Dulac, C. (2010). Brain function and chromatin plasticity. *Nature*, 465(7299), 728–735.
- Dunbar, K. N. (2008). *Learning, arts, and the brain*. The Dana consortium report on Arts and Cognition. Dunbar section Arts, Education, The Brain and language, pp. 81–92.
- Egan, M. F., Goldberg, T. E., Kolachana, B. S., Callicott, J. H., Mazzanti, C. M., Straub, R. E., et al. (2001). Effect of COMT Val108/158 Met genotype on frontal lobe function and risk for schizophrenia. *Proceedings of the National Academy of Sciences USA*, 98(12), 6917–6922.
- Egan, M. F., Kojima, M., Callicott, J. H., Goldberg, T. E., Kolachana, B. S., Bertolino, A., et al. (2003). The BDNF val66met polymorphism affects activity-dependent secretion of BDNF and human memory and hippocampal function. *Cell*, 112(2), 257–269.
- Egan, M. F., Straub, R. E., Goldberg, T. E., Yakub, I., Callicott, J. H., Hariri, A. R., et al. (2004). Variation in GRM3 affects cognition, prefrontal glutamate, and risk for schizophrenia. *Proceedings of the National Academy of Sciences USA*, 101(34), 12604–12609.
- Emery, M. R., Mattay, V. S., Sambataro, F., Lemaitre, H. S., Goldman, A. L., Tan, H. Y., et al. (2009). WWC1 (KIBRA) genotype modulates hippocampal structure and episodic memory-related neural activity. *Neuroimage*, 47(Suppl 1), S54.
- Enard, W., Przeworski, M., Fisher, S. E., Lai, C. S., Wiebe, V., Kitano, T., et al. (2002). Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*, 418(6900), 869–872.
- Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, 11(1), 19–23.
- Espeseth, T., Greenwood, P. M., Reinvang, I., Fjell, A. M., Walhovd, K. B., Westlye, L. T., et al. (2006). Interactive effects of APOE and CHRNA4 on attention and white matter volume in healthy middle-aged and older adults. *Cognitive, Affective, and Behavioral Neuroscience*, 6(1), 31–43.
- Fallgatter, A. J., Herrmann, M. J., Roemmler, J., Ehlis, A. C., Wagener, A., Heidrich, A., et al. (2004). Allelic variation of serotonin transporter function modulates the brain electrical response for error processing. *Neuropsychopharmacology*, 29(8), 1506–1511.
- Fan, J., Fossella, J., Sommer, T., Wu, Y., & Posner, M. I. (2003). Mapping the genetic variation of executive attention onto brain activity. *Proceedings of the National Academy of Sciences USA*, 100(12), 7406–7411.
- Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience*, 14(3), 340–347.
- Fan, J., & Posner, M. (2004). Human attentional networks. *Psychiatrice Praxis*, 31(Suppl 2), S210–S214.
- Fan, J., Wu, Y., Fossella, J. A., & Posner, M. I. (2001). Assessing the heritability of attentional networks. *BMC Neuroscience*, 2, 14.
- Filbey, F. M., Slack, K. J., Sunderland, T. P., & Cohen, R. M. (2006). Functional magnetic resonance imaging and magnetoencephalography differences associated with APOEepsilon4 in young healthy adults. *Neuroreport*, 17(15), 1585–1590.
- Fisher, S. E. (2006). Tangled webs: Tracing the connections between genes and cognition. *Cognition*, 101(2), 270–297.
- Fisher, S. E., & DeFries, J. C. (2002). Developmental dyslexia: Genetic dissection of a complex cognitive trait. *Nature Reviews Neuroscience*, 3(10), 767–780.
- Fisher, S. E., Lai, C. S., & Monaco, A. P. (2003). Deciphering the genetic basis of speech and language disorders. *Annual Review of Neuroscience*, 26, 57–80.
- Flint, J., & Munafò, M. R. (2007). The endophenotype concept in psychiatric genetics. *Psychol Med*, 37(2), 163–180.
- Fossella, J., & Posner, M. I. (2004). Genes and the development of neural networks underlying cognitive processes. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences III* (3 ed., pp. 1255–1266). Cambridge, MA: MIT Press.
- Fossella, J., Green, A. E., & Fan, J. (2006). Evaluation of a structural polymorphism in the ankyrin repeat and kinase domain containing 1 (ANKK1) gene and the activation of executive attention networks. *Cognitive, Affective, and Behavioral Neuroscience*, 6(1), 71–78.
- Fossella, J., Sommer, T., Fan, J., Wu, Y., Swanson, J. M., Pfaff, D. W., et al. (2002). Assessing the molecular-genetics of attention networks. *BMC Neuroscience*, 3, 14.
- Fraga, M. F., Ballestar, E., Paz, M. F., Ropero, S., Setien, F., Ballestar, M. L., et al. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences USA*, 102(30), 10604–10609.
- Gabrieli, J. D. (1998). Cognitive neuroscience of human memory. *Annual Review of Psychology*, 49, 87–115.
- Gallinat, J., Gotz, T., Kalus, P., Bajbouj, M., Sander, T., & Winterer, G. (2007). Genetic variations of the NR3A subunit of the NMDA receptor modulate prefrontal cerebral activity in humans. *Journal of Cognitive Neuroscience*, 19(1), 59–68.
- Goldberg, T. E., Egan, M. F., Gscheidle, T., Coppola, R., Weickert, T., Kolachana, B. S., et al. (2003). Executive subprocesses in working memory: Relationship to catechol-O-methyltransferase Val158Met genotype and schizophrenia. *Archives of General Psychiatry*, 60(9), 889–896.
- Goldberg, T. E., & Weinberger, D. R. (2004). Genes and the parsing of cognitive processes. *Trends in Cognitive Science*, 8(7), 325–335.
- Goodman, L. J., Valverde, J., Lim, F., Geschwind, M. D., Federoff, H. J., Geller, A. I., et al. (1996). Regulated release and polarized localization of brain-derived neurotrophic factor in hippocampal neurons. *Molecular and Cellular Neuroscience*, 7(3), 222–238.

- Graff, J., & Mansuy, I. M. (2008). Epigenetic codes in cognition and behaviour. *Behavioural Brain Research*, 192(1), 70–87.
- Graff, J., & Mansuy, I. M. (2009). Epigenetic dysregulation in cognitive disorders. *European Journal of Neuroscience*, 30(1), 1–8.
- Gray, J. R., Burgess, G. C., Schaefer, A., Yarkoni, T., Larsen, R. J., & Braver, T. S. (2005). Affective personality differences in neural processing efficiency confirmed using fMRI. *Cognitive, Affective, and Behavioral Neuroscience*, 5(2), 182–190.
- Gray, J. R., Chabris, C. F., & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience*, 6(3), 316–322.
- Gray, J. R., & Thompson, P. M. (2004). Neurobiology of intelligence: Science and ethics. *Nature Reviews Neuroscience*, 5(6), 471–482.
- Green, A. E., Kraemer, D. J., Fugelsang, J., Gray, J. R., & Dunbar, K. (2010). Connecting long distance: Semantic distance in analogical reasoning modulates frontopolar cortex activity. *Cereb Cortex*, 20, 70–76.
- Green, A. E., Kraemer, D. J., Fugelsang, J., Gray, J. R., & Dunbar, K. (in press). Journal of Experimental Psychology: Learning, Memory, and Cognition.
- Green, A. E., Kraemer, D. J., DeYoung, C. G., Fossella, J., & Gray, J. R. (submitted). A Gene-Brain-Cognition Pathway for the Effect of COMT on Executive Attention and IQ.
- Green, A. E., Munafo, M. R., DeYoung, C. G., Fossella, J. A., Fan, J., & Gray, J. R. (2008). Using genetic data in cognitive neuroscience: from growing pains to genuine insights. *Nature Reviews Neuroscience*, 9(9), 710–720.
- Greenwood, P. M., Lin, M. K., Sundararajan, R., Fryxell, K. J., & Parasuraman, R. (2009). Synergistic effects of genetic variation in nicotinic and muscarinic receptors on visual attention but not working memory. *Proceedings of the National Academy of Sciences USA*, 106(9), 3633–3638.
- Hariri, A. R., Goldberg, T. E., Mattay, V. S., Kolachana, B. S., Callicott, J. H., Egan, M. F., et al. (2003). Brain-derived neurotrophic factor val66met polymorphism affects human memory-related hippocampal activity and predicts memory performance. *Journal of Neuroscience*, 23(17), 6690–6694.
- Hashimoto, R., Moriguchi, Y., Yamashita, F., Mori, T., Nemoto, K., Okada, T., et al. (2008). Dose-dependent effect of the Val66Met polymorphism of the brain-derived neurotrophic factor gene on memory-related hippocampal activity. *Neuroscience Research*, 61(4), 360–367.
- Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5), 393–402.
- Ho, B. C., Milev, P., O'Leary, D. S., Librant, A., Andreasen, N. C., & Wassink, T. H. (2006). Cognitive and magnetic resonance imaging brain morphometric correlates of brain-derived neurotrophic factor Val66Met gene polymorphism in patients with schizophrenia and healthy volunteers. *Archives of General Psychiatry*, 63(7), 731–740.
- Huentelman, M. J., Papassotiropoulos, A., Craig, D. W., Hoerndli, F. J., Pearson, J. V., Huynh, K. D., et al. (2007). Calmodulin-binding transcription activator 1 (CAMTA1) alleles predispose human episodic memory performance. *Human Molecular Genetics*, 16(12), 1469–1477.
- Jansen, A., Krach, S., Krug, A., Markov, V., Eggemann, T., Zerres, K., et al. (2009). A putative high risk diplotype of the G72 gene is in healthy individuals associated with better performance in working memory functions and altered brain activity in the medial temporal lobe. *Neuroimage*, 45(3), 1002–1008.
- Joober, R., Gauthier, J., Lal, S., Bloom, D., Lalonde, P., Rouleau, G., et al. (2002). Catechol-O-methyltransferase Val-108/158-Met gene variants associated with performance on the Wisconsin Card Sorting Test. *Archives of General Psychiatry*, 59(7), 662–663.
- Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99(1), 122–149.
- Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin Review*, 9(4), 637–671.
- Kang, H., & Schuman, E. M. (1995). Long-lasting neurotrophin-induced enhancement of synaptic transmission in the adult hippocampus. *Science*, 267(5204), 1658–1662.
- Kircher, T., Krug, A., Markov, V., Whitney, C., Krach, S., Zerres, K., et al. (2009). Genetic variation in the schizophrenia-risk gene neuregulin 1 correlates with brain activation and impaired speech production in a verbal fluency task in healthy individuals. *Human Brain Mapping*, 30(10), 3406–3416.
- Koten, J. W., Jr., Wood, G., Hagoort, P., Goebel, R., Propsting, P., Willmes, K., et al. (2009). Genetic contribution to variation in cognitive function: An fMRI study in twins. *Science*, 323(5922), 1737–1740.
- Kovas, Y., & Plomin, R. (2006). Generalist genes: Implications for the cognitive sciences. *Trends in Cognitive Science*, 10(5), 198–203.
- Krabbendam, L., Isusi, P., Galdos, P., Echevarria, E., Bilbao, J. R., Martin-Pagola, A., et al. (2006). Associations between COMTVal158Met polymorphism and cognition: Direct or indirect effects? *European Psychiatry*, 21(5), 338–342.
- Krach, S., Jansen, A., Krug, A., Markov, V., Thimm, M., Sheldrick, A. J., et al. (2010). COMT genotype and its role on hippocampal-prefrontal regions in declarative memory. *Neuroimage*, 53, 978–984.
- Krug, A., Markov, V., Sheldrick, A., Krach, S., Jansen, A., Zerres, K., et al. (2009). The effect of the COMT val(158)met polymorphism on neural correlates of semantic verbal fluency. *European Archives of Psychiatry and Clinical Neuroscience*, 259(8), 459–465.
- Krug, A., Nieratschker, V., Markov, V., Krach, S., Jansen, A., Zerres, K., et al. (2010). Effect of CACNA1C rs1006737 on neural correlates of verbal fluency in healthy individuals. *Neuroimage*, 49(2), 1831–1836.
- Lai, C. S., Fisher, S. E., Hurst, J. A., Vargha-Khadem, F., & Monaco, A. P. (2001). A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature*, 413(6855), 519–523.
- Lai, C. S., Gerrelli, D., Monaco, A. P., Fisher, S. E., & Copp, A. J. (2003). FOXP2 expression during brain development coincides with adult sites of pathology in a severe speech and language disorder. *Brain*, 126(Pt 11), 2455–2462.
- Lamprecht, R., & LeDoux, J. (2004). Structural plasticity and memory. *Nature Reviews Neuroscience*, 5(1), 45–54.
- Le Hellard, S., Havik, B., Espeseth, T., Breilid, H., Lovlie, R., Luciano, M., et al. (2009). Variants in doublecortin- and calmodulin kinase like 1, a gene up-regulated by BDNF, are associated with memory and general cognitive abilities. *PLoS One*, 4(10), e7534.
- Lewis, D. A., Melchitzky, D. S., Sesack, S. R., Whitehead, R. E., Auh, S., & Sampson, A. (2001). Dopamine transporter immunoreactivity in monkey cerebral cortex: Regional,

- laminar, and ultrastructural localization. *Journal of Comparative Neurology*, 432(1), 119–136.
- Liegeois, F., Baldeweg, T., Connelly, A., Gadian, D. G., Mishkin, M., & Vargha-Khadem, F. (2003). Language fMRI abnormalities associated with FOXP2 gene mutation. *Nature Neuroscience*, 6(11), 1230–1237.
- Lotta, T., Vidgren, J., Tilgmann, C., Ulmanen, I., Melen, K., Julkunen, I., et al. (1995). Kinetics of human soluble and membrane-bound catechol O-methyltransferase: A revised mechanism and description of the thermolabile variant of the enzyme. *Biochemistry*, 34(13), 4202–4210.
- Malhotra, A. K., Kestler, L. J., Mazzanti, C., Bates, J. A., Goldberg, T., & Goldman, D. (2002). A functional polymorphism in the COMT gene and performance on a test of prefrontal cognition. *American Journal of Psychiatry*, 159(4), 652–654.
- Marenci, S., Steele, S. U., Egan, M. F., Goldberg, T. E., Straub, R. E., Sharrief, A. Z., et al. (2006). Effect of metabotropic glutamate receptor 3 genotype on N-acetylaspartate measures in the dorsolateral prefrontal cortex. *American Journal of Psychiatry*, 163(4), 740–742.
- Markov, V., Krug, A., Krach, S., Whitney, C., Eggermann, T., Zerres, K., et al. (2009). Genetic variation in schizophrenia-risk-gene dysbindin 1 modulates brain activation in anterior cingulate cortex and right temporal gyrus during language production in healthy individuals. *Neuroimage*, 47(4), 2016–2022.
- Masterpasqua, F. (2009). Psychology and epigenetics. *Review of General Psychology*, 13, 194–201.
- Matsumoto, M., Weickert, C. S., Akil, M., Lipska, B. K., Hyde, T. M., Herman, M. M., et al. (2003). Catechol O-methyltransferase mRNA expression in human and rat brain: Evidence for a role in cortical neuronal function. *Neuroscience*, 116(1), 127–137.
- Mattay, V. S., Goldberg, T. E., Fera, F., Hariri, A. R., Tessitore, A., Egan, M. F., et al. (2003). Catechol O-methyltransferase val158-met genotype and individual variation in the brain response to amphetamine. *Proceedings of the National Academy of Sciences USA*, 100(10), 6186–6191.
- McClearn, G. E., Johansson, B., Berg, S., Pedersen, N. L., Ahern, F., Petrill, S. A., et al. (1997). Substantial genetic influence on cognitive abilities in twins 80 or more years old. *Science*, 276(5318), 1560–1563.
- McGowan, P., & Szyf, M. (2010). Environmental epigenomics: Understanding the effect of parental care on the epigenome. *Essays in Biochemistry*, 48, 275–287.
- Mesulam, M. M. (1981). A cortical network for directed attention and unilateral neglect. *Annals of Neurology*, 10(4), 309–325.
- Meyer-Lindenberg, A., Buckholtz, J. W., Kolachana, B., A, R. H., Pezawas, L., Blasi, G., et al. (2006). Neural mechanisms of genetic risk for impulsivity and violence in humans. *Proceedings of the National Academy of Sciences USA*, 103(16), 6269–6274.
- Meyer-Lindenberg, A., Kohn, P. D., Kolachana, B., Kippenhan, S., McInerney-Leo, A., Nussbaum, R., et al. (2005). Midbrain dopamine and prefrontal function in humans: Interaction and modulation by COMT genotype. *Nature Neuroscience*, 8(5), 594–596.
- Meyer-Lindenberg, A., Nichols, T., Callicott, J. H., Ding, J., Kolachana, B., Buckholtz, J., et al. (2006). Impact of complex genetic variation in COMT on human brain function. *Molecular Psychiatry*, 11(9), 867–877, 797.
- Meyer-Lindenberg, A., Nicodemus, K. K., Egan, M. F., Callicott, J. H., Mattay, V., & Weinberger, D. R. (2008). False positives in imaging genetics. *Neuroimage*, 40, 655–661.
- Meyer-Lindenberg, A., & Weinberger, D. R. (2006). Intermediate phenotypes and genetic mechanisms of psychiatric disorders. *Nature Reviews Neuroscience*, 7(10), 818–827.
- Moosmang, S., Haider, N., Klugbauer, N., Adelsberger, H., Langwieser, N., Muller, J., et al. (2005). Role of hippocampal Cav1.2 Ca²⁺ channels in NMDA receptor-independent synaptic plasticity and spatial memory. *Journal of Neuroscience*, 25(43), 9883–9892.
- Munafo, M. R. (2006). Candidate gene studies in the 21st century: Meta-analysis, mediation, moderation. *Genes, Brain and Behavior*, 5(Suppl 1), 3–8.
- Murer, M. G., Yan, Q., & Raisman-Vozari, R. (2001). Brain-derived neurotrophic factor in the control human brain, and in Alzheimer's disease and Parkinson's disease. *Progress in Neurobiology*, 63(1), 71–124.
- Nackley, A. G., Shabalina, S. A., Tchivileva, I. E., Satterfield, K., Korchynskyi, O., Makarov, S. S., et al. (2006). Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science*, 314(5807), 1930–1933.
- Nagel, I. E., Chicherio, C., Li, S. C., von Oertzen, T., Sander, T., Villringer, A., et al. (2008). Human aging magnifies genetic effects on executive functioning and working memory. *Frontiers in Human Neuroscience*, 2, 1.
- Nave, K. A., & Salzer, J. L. (2006). Axonal regulation of myelination by neuregulin 1. *Current Opinion in Neurobiology*, 16(5), 492–500.
- Nystrom, L. E., Braver, T. S., Sabb, F. W., Delgado, M. R., Noll, D. C., & Cohen, J. D. (2000). Working memory for letters, shapes, and locations: fMRI evidence against stimulus-based regional organization in human prefrontal cortex. *Neuroimage*, 11(5, Pt 1), 424–446.
- Otnaess, M. K., Djurovic, S., Rimol, L. M., Kulle, B., Kahler, A. K., Jonsson, E. G., et al. (2009). Evidence for a possible association of neurotrophin receptor (NTRK-3) gene polymorphisms with hippocampal function and schizophrenia. *Neurobiology of Disease*, 34(3), 518–524.
- Papassotiropoulos, A., Henke, K., Aerni, A., Coluccia, D., Garcia, E., Wollmer, M. A., et al. (2005). Age-dependent effects of the 5-hydroxytryptamine-2a-receptor polymorphism (His452Tyr) on human memory. *Neuroreport*, 16(8), 839–842.
- Papassotiropoulos, A., Stephan, D. A., Huentelman, M. J., Hoerndl, F. J., Craig, D. W., Pearson, J. V., et al. (2006). Common Kibra alleles are associated with human memory performance. *Science*, 314(5798), 475–478.
- Papassotiropoulos, A., Wollmer, M. A., Aguzzi, A., Hock, C., Nitsch, R. M., & de Quervain, D. J. (2005). The prion gene is associated with human long-term memory. *Human Molecular Genetics*, 14(15), 2241–2246.
- Parasuraman, R., Greenwood, P. M., Kumar, R., & Fossella, J. (2005). Beyond heritability: Neurotransmitter genes differentially modulate visuospatial attention and working memory. *Psychological Science*, 16(3), 200–207.
- Passamonti, L., Cerasa, A., Gioia, M. C., Magariello, A., Muglia, M., Quattrone, A., et al. (2008). Genetically dependent modulation of serotonergic inactivation in the human prefrontal cortex. *Neuroimage*, 40(3), 1264–1273.
- Passamonti, L., Fera, F., Magariello, A., Cerasa, A., Gioia, M. C., Muglia, M., et al. (2006). Monoamine oxidase-a genetic variations influence brain activity associated with inhibitory control: New insight into the neural correlates of impulsivity. *Biological Psychiatry*, 59(4), 334–340.

- Pezawas, L., Verchinski, B. A., Mattay, V. S., Callicott, J. H., Kolachana, B. S., Straub, R. E., et al. (2004). The brain-derived neurotrophic factor val66met polymorphism and variation in human cortical morphology. *Journal of Neuroscience*, 24(45), 10099–10102.
- Plomin, R. (2000). Psychology in a post-genomics world: It will be more important than ever. *Observer: American Psychological Society*, 13, 3.
- Poo, M. M. (2001). Neurotrophins as synaptic modulators. *Nature Reviews Neuroscience*, 2(1), 24–32.
- Porjesz, B., Almasy, L., Edenberg, H. J., Wang, K., Chorlian, D. B., Foroud, T., et al. (2002). Linkage disequilibrium between the beta frequency of the human EEG and a GABA_A receptor gene locus. *Proceedings of the National Academy of Sciences USA*, 99(6), 3729–3733.
- Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, 13, 25–42.
- Posner, M. I., Rothbart, M. K., & Sheese, B. E. (2007). Attention genes. *Developmental Science*, 10(1), 24–29.
- Potkin, S. G., Turner, J. A., Fallon, J. A., Lakatos, A., Keator, D. B., Guffanti, G., et al. (2009). Gene discovery through imaging genetics: Identification of two novel genes associated with schizophrenia. *Molecular Psychiatry*, 14(4), 416–428.
- Qiu, J. (2006). Epigenetics: Unfinished symphony. *Nature*, 441, 143–145.
- Ramnani, N., & Owen, A. M. (2004). Anterior prefrontal cortex: Insights into function from anatomy and neuroimaging. *Nature Reviews Neuroscience*, 5(3), 184–194.
- Rasch, B., Papassotiropoulos, A., & de Quervain, D. F. (2010). Imaging genetics of cognitive functions: Focus on episodic memory. *Neuroimage*, 53, 870–877.
- Rasch, B., Spalek, K., Buhholzer, S., Luechinger, R., Boesiger, P., Papassotiropoulos, A., et al. (2009). A genetic variation of the noradrenergic system is related to differential amygdala activation during encoding of emotional memories. *Proceedings of the National Academy of Sciences USA*, 106(45), 19191–19196.
- Raz, A., & Buhle, J. (2006). Typologies of attentional networks. *Nature Reviews Neuroscience*, 7(5), 367–379.
- Relton, C., & Smith, G. D. (2010). Epigenetic epidemiology of common complex disease: Prospects for prediction, prevention, and treatment. *PLoS Med*, 7, 1–8.
- Reuter, M., Esslinger, C., Montag, C., Lis, S., Gallhofer, B., & Kirsch, P. (2008). A functional variant of the tryptophan hydroxylase 2 gene impacts working memory: A genetic imaging study. *Biological Psychology*, 79(1), 111–117.
- Reuter, M., Ott, U., Vaitl, D., & Hennig, J. (2007). Impaired executive control is associated with a variation in the promoter region of the tryptophan hydroxylase 2 gene. *Journal of Cognitive Neuroscience*, 19(3), 401–408.
- Reuter, M., Roth, S., Holte, K., & Hennig, J. (2006). Identification of first candidate genes for creativity: A pilot study. *Brain Research*, 1069(1), 190–197.
- Riccio, A. (2010). Dynamic epigenetic regulation in neurons: Enzymes, stimuli and signaling pathways. *Nature Neuroscience*, 13(11), 1330–1337.
- Rotaru, D. C., Lewis, D. A., & Gonzalez-Burgos, G. (2007). Dopamine D1 receptor activation regulates sodium channel-dependent EPSP amplification in rat prefrontal cortex pyramidal neurons. *Journal of Physiology*, 581(Pt 3), 981–1000.
- Rueda, M. R., Rothbart, M. K., McCandliss, B. D., Saccomanno, L., & Posner, M. I. (2005). Training, maturation, and genetic influences on the development of executive attention. *Proceedings of the National Academy of Sciences USA*, 102(41), 14931–14936.
- Rujescu, D., Meisenzahl, E. M., Krejcová, S., Giegling, I., Zetsche, T., Reiser, M., et al. (2007). Plexin B3 is genetically associated with verbal performance and white matter volume in human brain. *Molecular Psychiatry*, 12(2), 190–194, 115.
- Rybakowski, J. K., Borkowska, A., Skibinska, M., & Hauser, J. (2006). Illness-specific association of val66met BDNF polymorphism with performance on Wisconsin Card Sorting Test in bipolar mood disorder. *Molecular Psychiatry*, 11(2), 122–124.
- Sambataro, F., Reed, J. D., Murty, V. P., Das, S., Tan, H. Y., Callicott, J. H., et al. (2009). Catechol-O-methyltransferase valine(158)methionine polymorphism modulates brain networks underlying working memory across adulthood. *Biological Psychiatry*, 66(6), 540–548.
- Sarter, M., & Bruno, J. P. (1997). Cognitive functions of cortical acetylcholine: Toward a unifying hypothesis. *Brain Research and Brain Research Review*, 23(1–2), 28–46.
- Savitz, J., Solms, M., & Ramesar, R. (2006). The molecular-genetics of cognition: Dopamine, COMT and BDNF. *Genes, Brain and Behavior*, 5(4), 311–328.
- Schofield, P. R., Williams, L. M., Paul, R. H., Gatt, J. M., Brown, K., Luty, A., et al. (2009). Disturbances in selective information processing associated with the BDNF Val66Met polymorphism: Evidence from cognition, the P300 and fronto-hippocampal systems. *Biological Psychology*, 80(2), 176–188.
- Schott, B. H., Seidenbecher, C. I., Fenker, D. B., Lauer, C. J., Bunzeck, N., Bernstein, H. G., et al. (2006). The dopaminergic midbrain participates in human episodic memory formation: Evidence from genetic imaging. *Journal of Neuroscience*, 26(5), 1407–1417.
- Silva, A. J. (2003). Molecular and cellular cognitive studies of the role of synaptic plasticity in memory. *Journal of Neurobiology*, 54(1), 224–237.
- Smith, E. E., & Jonides, J. (1998). Neuroimaging analyses of human working memory. *Proceedings of the National Academy of Sciences USA*, 95(20), 12061–12068.
- Stollstorf, M., Foss-Feig, J., Cook, E. H., Jr., Stein, M. A., Gaillard, W. D., & Vaidya, C. J. (2010). Neural response to working memory load varies by dopamine transporter genotype in children. *Neuroimage*, 53, 970–977.
- Szeszko, P. R., Lipsky, R., Mentschel, C., Robinson, D., Gunduz-Bruce, H., Sevy, S., et al. (2005). Brain-derived neurotrophic factor val66met polymorphism and volume of the hippocampal formation. *Molecular Psychiatry*, 10(7), 631–636.
- Tan, H. Y., Callicott, J. H., & Weinberger, D. R. (2008). Intermediate phenotypes in schizophrenia genetics redux: Is it a no brainer? *Molecular Psychiatry*, 13(3), 233–238.
- Tan, H. Y., Chen, Q., Sust, S., Buckholtz, J. W., Meyers, J. D., Egan, M. F., et al. (2007). Epistasis between catechol-O-methyltransferase and type II metabotropic glutamate receptor 3 genes on working memory brain function. *Proceedings of the National Academy of Sciences USA*, 104(30), 12536–12541.
- Tan, H. Y., Nicodemus, K. K., Chen, Q., Li, Z., Brooke, J. K., Honea, R., et al. (2008). Genetic variation in AKT1 is linked to dopamine-associated prefrontal cortical structure and function in humans. *Journal of Clinical Investigation*, 118(6), 2200–2208.
- Thompson, P. M., Cannon, T. D., Narr, K. L., van Erp, T., Poutanen, V. P., Huttunen, M., et al. (2001). Genetic influences on brain structure. *Nature Neuroscience*, 4(12), 1253–1258.

- Toro, R., Chupin, M., Garnero, L., Leonard, G., Perron, M., Pike, B., et al. (2009). Brain volumes and Val66Met polymorphism of the BDNF gene: Local or global effects? *Brain Struct Funct*, 213(6), 501–509.
- Tsai, S. J., Hong, C. J., Liu, M. E., Hou, S. J., Yen, F. C., Hsieh, C. H., & Liou, Y. J. (2008). Interleukin-1 beta (C-511T) genetic polymorphism is associated with cognitive performance in elderly males without dementia. *Neurobiology of Aging*, 31, 1950–1955.
- Tsai, S. J., Yu, Y. W., Chen, T. J., Chen, J. Y., Liou, Y. J., Chen, M. C., et al. (2003). Association study of a functional catechol-O-methyltransferase-gene polymorphism and cognitive function in healthy females. *Neuroscience Letters*, 338(2), 123–126.
- Tunbridge, E. M., Weickert, C. S., Kleinman, J. E., Herman, M. M., Chen, J., Kolachana, B. S., et al. (2007). Catechol-o-methyltransferase enzyme activity and protein expression in human prefrontal cortex across the postnatal lifespan. *Cerebral Cortex*, 17(5), 1206–1212.
- Turkheimer, E., Haley, A., Waldron, M., D'Onofrio, B., & Gottesman, II. (2003). Socioeconomic status modifies heritability of IQ in young children. *Psychological Science*, 14(6), 623–628.
- Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent. *Journal of Memory and Language*, 28, 127–154.
- Voelker, P., Sheese, B. E., Rothbart, M. K., & Posner, M. I. (2009). Variations in catechol-O-methyltransferase gene interact with parenting to influence attention in early development. *Neuroscience*, 164(1), 121–130.
- Vogler, C., Spalek, K., Aerni, A., Demougin, P., Muller, A., Huynh, K. D., et al. (2009). CPEB3 is associated with human episodic memory. *Frontiers in Behavioral Neuroscience*, 3, 4.
- Waldman, I. D., Nigg, J. T., Gizer, I. R., Park, L., Rapley, M. D., & Friderici, K. (2006). The adrenergic receptor alpha-2A gene (ADRA2A) and neuropsychological executive functions as putative endophenotypes for childhood ADHD. *Cognitive, Affective, and Behavioral Neuroscience*, 6(1), 18–30.
- Watkins, K. E., Vargha-Khadem, F., Ashburner, J., Passingham, R. E., Connelly, A., Friston, K. J., et al. (2002). MRI analysis of an inherited speech and language disorder: structural brain abnormalities. *Brain*, 125(Pt 3), 465–478.
- Weinberger, D. R., Egan, M. F., Bertolino, A., Callicott, J. H., Mattay, V. S., Lipska, B. K., et al. (2001). Prefrontal neurons and the genetics of schizophrenia. *Biological Psychiatry*, 50(11), 825–844.
- White, J. A., McKinney, B. C., John, M. C., Powers, P. A., Kamp, T. J., & Murphy, G. G. (2008). Conditional forebrain deletion of the L-type calcium channel Ca V 1.2 disrupts remote spatial memories in mice. *Learning and Memory*, 15(1), 1–5.
- Winterer, G., Musso, F., Konrad, A., Vucurevic, G., Stoeter, P., Sander, T., et al. (2007). Association of attentional network function with exon 5 variations of the CHRNA4 gene. *Human Molecular Genetics*, 16(18), 2165–2174.
- Winterer, G., Musso, F., Vucurevic, G., Stoeter, P., Konrad, A., Seker, B., et al. (2006). COMT genotype predicts BOLD signal and noise characteristics in prefrontal circuits. *Neuroimage*, 32(4), 1722–1732.
- Winterer, G., & Weinberger, D. R. (2004). Genes, dopamine and cortical signal-to-noise ratio in schizophrenia. *Trends in Neuroscience*, 27(11), 683–690.
- Wishart, H. A., Saykin, A. J., Rabin, L. A., Santulli, R. B., Flashman, L. A., Guerin, S. J., et al. (2006). Increased brain activation during working memory in cognitively intact adults with the APOE epsilon4 allele. *American Journal of Psychiatry*, 163(9), 1603–1610.
- Wolf, C., Jackson, M. C., Kissling, C., Thome, J., & Linden, D. E. (2009). Dysbindin-1 genotype effects on emotional working memory. *Molecular Psychiatry*, 16, 145–155.

This page intentionally left blank

PART

2

Deductive, Inductive,
and Abductive
Reasoning

This page intentionally left blank

Dual-Process Theories of Deductive Reasoning: Facts and Fallacies

Jonathan St. B. T. Evans

Abstract

The psychology of reasoning was dominated by the deduction paradigm from around 1960 to 2000, in which untrained participants are asked to assess the validity of logical arguments. As evidence of logical error and content-dependent thinking amassed, the paradigm has shifted recently with more emphasis on probabilistic and pragmatic processes. This chapter is focused particularly on the dual-process theories that arose from traditional studies of deductive reasoning but that now form part of a more general set of theories of higher cognition. It is argued that the “received” view of dual-process theory, which was established around 2000, actually incorporates a number of false beliefs and fallacies, which are discussed in this chapter. While dual-process theory rightly remains the focus of much current research, it is important to understand and avoid these fallacies.

Key Words: deductive reasoning, dual-process theory, inference

Introduction

The idea that there are two fundamentally different kinds of thinking has been proposed for centuries by numerous philosophers and psychologists. In many cases, such dual-process theories have been developed in ignorance of the similar ideas of both earlier and contemporary authors (Frankish & Evans, 2009). In postwar psychology, the earliest dual-process theory seems to have been that developed by Reber (1993) founded in his long program of research into implicit and explicit learning, which started in the 1960s. Reber was probably the first to suggest that implicit and explicit processes were derived from two distinct cognitive systems, one ancient and animal-like and the other recently evolved and distinctively human, an idea echoed in much later writing (Epstein, 1994; Evans, 2010; Evans & Over, 1996; Stanovich, 1999; Stanovich, 2004). Numerous dual-process theories of social cognition were also proposed in the early 1980s, including both the ideas that (a) social knowledge

may take an implicit form (e.g., stereotypes) that influences behavior independently of consciously accessible beliefs, and (b) information may be processed in an automatic or shallow manner, or in a controlled and effortful way (for recent reviews, see Deutsch & Strack, 2006; Lieberman, 2007; Smith & DeCoster, 2000, Smith & Collins, 2009).

In this chapter, I focus mostly on the theories that have arisen from research in the study of deductive reasoning, although the fallacies I shall identify may be seen in other fields in which dual-process theories are discussed. The psychology of deductive reasoning is a major tradition that dates from the early 20th century and which was intensively studied from the 1960s to date. It was built around the deduction paradigm (see Evans, 2002) in which participants are asked to assume the truth of some premises and decide whether a conclusion follows necessarily from them. Normally, participants are used who have no training in logic and none is provided by the experimenter. This paradigm derives from

a logicist tradition in philosophy and psychology which assumed that logic must be built into the mind in order to allow rational thought to take place (Henle, 1962; Inhelder & Piaget, 1958). As evidence of logical error and content-dependent thinking amassed, however, the foundations of the paradigm were gradually undermined, to the point where a distinct shift to a new paradigm psychology of reasoning appears to have occurred in the past 10–15 years. The inheritors of the deductive reasoning tradition now use a range of methods and draw on pragmatic and probabilistic accounts rather than logical processing (see Evans, *in press* for an account of the paradigm shift; also Oaksford & Chater, 2010 for many recent examples of studies based on the new paradigm). However, the deduction paradigm is still used, though interpreted somewhat differently, as we shall see later.

Dual-process theories in the deductive reasoning field have been around since the 1970s and until recently were developed independently and in ignorance of dual-process theories in social psychology and other fields. Nevertheless, the features typically attributed to the two kinds of thinking, which I shall call type 1 and type 2 in this chapter, are very similar to those proposed by authors in other fields (Evans, 2008). In Table 8.1, I list a set of the typical attributes in what I am going to call the “received” view of dual-process theory, based broadly on the writing of various authors during the 1990s. All dual-process theories seem

to contrast the idea of fast, automatic, low-effort and high-capacity processes (type 1) with slow, controlled, high-effort and low-capacity processes (type 2). The idea that type 1 thinking is unconscious and type 2 thinking is conscious has been a recurring theme in many accounts from Reber onward. The distinction between conscious and nonconscious processing has been particularly emphasized by social psychologists (e.g. Wilson, 2002) and was a feature of the earlier dual-process theory of reasoning (Wason & Evans, 1975). It has also been suggested that type 1 thinking operates preconsciously so that it precedes, shapes, and biases conscious type 2 thinking (e.g. Evans, 1989). The relationship of consciousness to dual-process theory is, however, rather more complex than implied by the received view. Consideration of this issue is beyond the scope of this chapter, but I have discussed it in detail elsewhere (Evans, 2010, Chapter 7).

The notion that type 2 reasoning involves explicit rule following, whereas type 1 processing relies on associative, experiential learning has been proposed by a number of authors writing across different fields of dual-process theory (Evans & Over, 1996; Sloman, 1996; Smith & DeCoster, 2000). Moreover, for historical reasons (see next section) there has been a strong association between type 1 processes and cognitive biases on the one hand, and type 2 processes with normatively correct reasoning on the other (see, for example, Evans, 1989; Stanovich, 1999). The same association is reflected in recent

Table 8.1. “Received” View of Dual-Process Theories of Reasoning and Higher Cognition, as It Emerged c. 2000

Type 1 Processes	Type 2 Processes
Unconscious, preconscious	Conscious
Rapid	Slow
Automatic	Controlled
Low effort	High effort
High capacity	Low capacity
Associative	Rule based
Intuitive	Deliberative
Contextualized	Abstract
Cognitive biases	Normative reasoning
Independent of cognitive capacity (IQ, WMC)	Correlated with individual differences in cognitive capacity
System 1	System 2

application of dual-process theory to the study of judgment and decision making by Kahneman and Frederick (2002). Stanovich (1999) particularly emphasized the distinction between contextually dependent type 1 processes and abstract, decontextualized type 2 thinking, talking of a *fundamental cognitive bias* to contextualize all information. In his empirical research program with West, Stanovich also demonstrated a strong association between cognitive ability (as measured mostly by SAT scores) and type 2 processing, while also showing that type 1 processes were broadly independent of such measures (Stanovich & West, 2000b).

Perhaps the most important feature of the received view of dual-process theory is that type 1 and 2 processes can be attributed to two distinct cognitive systems, now widely known as System 1 and 2 following the use of these terms by Stanovich (1999). This idea actually emerged from a combination of proposals made by several earlier authors (Epstein, 1994; Evans & Over, 1996; Reber, 1993). The two-system theory added a number of other features to the list shown in Table 8.1, in that System 1 was proposed to be an old system, which evolved early and shares many of its features with animal cognition, whereas System 2 was proposed to be recently evolved and either unique to humans or very distinctively developed in our own species. I have recently developed these concepts within a two-minds (rather than two-system) theory (Evans, 2010), but I will not discuss them in this chapter because the focus is on the more restricted dual-process theory. I will, however, discuss the idea that type 1 thinking can be allocated to just one system and type 2 thinking to another.

Having explained the received view of dual-process theory, I must now comment that both experimental research and theoretical analysis have moved on a lot in the past decade in such a way that it now appears many of the attributes listed in Table 8.1 are in some

ways suspect if not outright fallacious. However, even among dual-process researchers, awareness of some of these issues is not as high as it might be. So in the later part of this chapter, I am going to discuss a number of fallacies that I believe to be associated with research on dual process (see Table 8.2). Note that I use the term “fallacy” rather loosely here: Some of the points in Table 8.2 refer to genuine inferential errors, whereas others are simply false beliefs. All result in mistaken interpretation or application of dual-process theory, however. I hasten to add that it is not my purpose to reject the fundamental type 1 and 2 distinction. So I will precede my discussion of the “fallacies” with two other sections. In the first I will sketch a brief history of the psychology of deductive reasoning and show how the dual-process theory developed. This is important both to understanding how contemporary research on dual-process theory is conducted, as well as explaining the origin of some of the fallacies. In the following section, I will present the experimental and neuroscientific evidence which I believe should compel us to take the type 1 and 2 distinction seriously. Finally, I shall discuss the fallacies that require us to think much more carefully about the way in which we discuss dual processes and conduct research upon them.

Deductive Reasoning and the Origins of the Dual-Process Theory of Reasoning

The earliest studies of deductive reasoning were published prior to World War II and used classical (Aristotelian) syllogisms (Wilkins, 1928; Woodworth & Sells, 1935). Such syllogisms are comprised of two premises and a conclusion that may or may not follow logically from the premises. These early studies provide a foretaste of what was to come in the modern study of the topic. Participants were shown to endorse many fallacies, that is to say that conclusions followed logically when they did not. They were also shown to be systematically

Table 8.2. Some Fallacies Associated With Dual-Process Theories of Thinking and Reasoning

-
1. All dual-process theories are essentially the same

 2. There are just two systems underlying type 1 and 2 processing

 3. Type 1 processes are responsible for cognitive biases; type 2 processes for normatively correct responding

 4. Type 1 processing is contextualized, whereas type 2 processing is abstract

 5. Fast processing indicates the use of a type 1 rather than type 2 process

biased by both the linguistic form of the syllogisms and the believability of the content (see Evans, Newstead, & Byrne, 1993 for a detailed review of research on syllogistic reasoning). Despite these early findings, the logicist ethos of the 1960s saw the launch of a large program of research employing the deduction paradigm.

While a U.S. tradition continued to focus mostly on people's ability to evaluate classical syllogisms, the British psychologist Peter Wason extended the paradigm with ingenious tasks of his own invention, the most famous of which is the four-card selection task (Wason, 1966). This task went beyond the deduction paradigm by involving hypothesis testing, but it still used formal logic as a normative reference. It has been studied by the same research community as those employing the deduction task proper and traditionally treated by reviewers as part of the same field (e.g. Evans, 2007a; Evans et al., 1993; Manktelow, 1999). The standard abstract task involves four cards lying on a table, on whose exposed sides might be the values

A D 3 7

Participants are told that each card has a capital letter on one side and a single-digit number on the other. They are also told that the following rule applies to these four cards and may be true or false:

If there is an A on one side of the card, then
there is a 3 on the other side of the card.

Their task is to identify those cards, and only those cards, which must be turned over in order to decide whether the rule is true or false. Given this task, most people choose the A card and a number of others choose the A and the 3. Wason argued that neither choice was logically correct: Participants should turn over the A and the 7. The reason is that only by discovering a card with an A on one side that did *not* have a 3 on the other could one prove the rule false, and only these cards could lead to such a discovery. Wason also discovered that the task can be easy to solve when framed with realistic content (see Wason & Johnson-Laird, 1972). Researchers became fascinated with both the difficulty of the abstract task and the apparent ease of certain realistic versions, with the result that study of human reasoning was dominated by the task for the next 25 years or so (see Evans & Over, 2004, Chapter 5, for a review).

From the 1970s onward, psychologists also started to study conditional inference, which uses

the deduction paradigm proper. The focus of interest has been on what philosophers term "elimination inferences" (conditional introduction inferences have been relatively neglected; see Over, Evans, & Elqayam, 2010). There are four of these as follows:

Modus Ponens (MP)	If p then q; p, therefore q
Denial of the Antecedent (DA)	If p then q, not-p, therefore not-q
Affirmation of the Consequent (AC)	If p then q, q, therefore p
Modus Tollens (MT)	If p then q, not-q therefore not-p

Interest in the study of conditional inference has gradually increased to the point where it is nowadays more commonly studied than the selection task, although not necessarily with deductive reasoning instructions. Of the four inferences, two (MP, MT) are logically valid and two (DA, AC) considered to be classical fallacies. Early studies soon established (a) that the fallacies are commonly endorsed (in common with findings on syllogistic reasoning) and (b) that while MP is endorsed with near 100% frequency, the other valid inference MT is far from obvious, with endorsement rates of around 60%–70%.

Explaining these findings became an issue of the major debate between two rival accounts of deductive reasoning which took place mainly in the 1980s. One tradition, known as "mental logic" was introduced to cognitive psychologists by Martin Braine (1978), although Jean Piaget's theory of formal operations was a precursor in developmental psychology (Inhelder & Piaget, 1958). According to this view, people have a set of logical rules built into their minds to permit reasoning, which proceeds as a kind of mental proof. Well-developed mental logic theories are presented by Braine and O'Brien (1998b) and Rips (1994). While standard logic has a rule for Modus Tollens, these authors suggested that mental logics do not, in order to account for the relative difficulty of the inference. Instead, MT could be derived by error-prone indirect reasoning procedures. Endorsement of the fallacies, DA and AC, was attributed to a mechanism outside of mental logic, namely that of pragmatic implicature. The rival theory, based on mental models, was launched by Johnson-Laird (1983; see also Johnson-Laird & Byrne, 1991). According to this theory, people do not require any logical rules in their heads to perform deductive reasoning. Instead, they operate a simple semantic principle: a deduction is valid if

there are no counterexamples to it. In this theory, logical possibilities are represented by mental models that may be fleshed out or eliminated as reasoning proceeds. People are deductively competent in principle but fallible in practice, as they have limited working memory capacity with which to consider models and are subject to a number of processing biases (see Johnson-Laird, Chapter 9, for a detailed account of research on mental model theory).

The mental model theory also gave an account of the basic phenomena of conditional inference (Johnson-Laird & Byrne, 2002). Initially, people consider explicitly only one mental model “*p & q*,” which allows MP but not MT to be easily drawn. To draw MT, people must “flesh out” models representing other possibilities, such as “not-*p* and not-*q*.” The endorsement of fallacies could result from superficial processing or biconditional representation in which the possibility “not-*p* and *q*” is excluded. It is interesting to note that although neither mental logic nor mental model theories were formulated as dual-process accounts, each makes proposals that can easily be mapped on to the type 1 and 2 distinction. Both proposed fast, low-effort (type 1) processes: immediate inference with built-in rules like MP, or formulation of and inference from initial mental models. Equally, both proposed effortful, error-prone (type 2) processes: indirect reasoning procedures or fleshing out of implicit mental models.

While mental model theory became a major paradigm in the 1980s and 1990s, this was also the heyday of the Wason selection task, especially in its realistic form. Theorists debated whether facilitatory versions reflected Darwinian algorithms for social exchange and hazard avoidance (Cosmides, 1989; Fiddick, Cosmides, & Tooby, 2000), acquired pragmatic reasoning schemas (Cheng & Holyoak, 1985; Holyoak & Cheng, 1995), or domain-general procedures based on either decision theoretic principles (Manktelow & Over, 1991; Oaksford & Chater, 1994) or pragmatic relevance (Sperber, Cara, & Girotto, 1995). In my view, these debates were never satisfactorily resolved. However, they did contribute to the downfall of the deduction paradigm in its traditional form, as all of these theorists emphasized the nonlogical nature of the cognitive processing involved.

The first explicit dual-process theory emerged in the 1970s in the context of the original abstract version of the selection task. Wason had conducted a number of experiments with the selection task

in which the erroneous patterns persisted and in which 10% or less of participants gave the correct solution. He found participants to be resistant to various “therapies” he devised to encourage insight into the task and concluded that they were fundamentally irrational, as he never doubted the logical standard for rationality (see Evans, 2002). However, this thinking was challenged when a very simple method of facilitating logical selections on the task was discovered. This involved putting a negation into the second part of the conditional statement. Suppose for the same set of cards shown earlier the rule was as follows:

If there is an A on one side of the card, then there is NOT a 3 on the other side of the card.

Now, the correct choice is A and 3 because a card with these values would falsify the claim. And this is what most participants also select, turning a very difficult problem into an easy one (first shown by Evans & Lynch, 1973; there have been many later replications as reviewed by Evans, 1998). The tendency to select cards named in the rule was termed “matching bias” and created a theoretical puzzle for Wason. He had already shown, in other research on the task, that participants gave verbal justifications for their choices that seemed to reveal the degree of insight that these choices suggested. But no one mentioned matching bias as a cause of his or her selections.

Wason and I collaborated on two papers to try to understand what was going on (Wason & Evans, 1975; Evans & Wason, 1976). In the first of these we gave both the hard affirmative and easy negative version to participants and asked them to write verbal justifications for their choices. We found, first of all, a strong matching bias on both tasks regardless of the order in which we presented them. We also found that participants always gave a justification consistent with their choices. So if they received the negative rule first and chose the correct (and matching) cards, they would say they were trying to falsify the statement. But when given the affirmative rule second, this apparent insight disappeared and with it the correct choices. Now they said they were trying to prove the statement true. And, of course, no one gave matching bias as the reason for his or her choices. Wason and Evans concluded that an unconscious type 1 process (matching) was responsible for the choices, while conscious type 2 processes were simply post hoc rationalizations, a conclusion supported by a different method in our second paper.

The original Wason and Evans study has also been replicated recently in essential respects by Lucas and Ball (2005) whose findings broadly supported the original conclusions.

A question left hanging by the 1975 paper was whether explicit reasoning (type 2 processing) did more than simply rationalize card choices on this task. Dual-process accounts typically propose a competition between two processes in determining the actual behavior observed. But for a long time it seemed as though card choices on the selection task might simply reflect the attention that was paid to individual cards. For example, using a computer-presented version of the task, Evans (1996) asked participants to point the mouse to cards they were thinking of selecting, thus measuring card inspection times. It was found that much more time was spent inspecting cards that were subsequently selected, suggesting that the “reasoning” observed was mostly to justify preconsciously cued selections. Although the mouse-pointing method was criticized (Roberts, 1998), all essential findings were replicated using eye-movement tracking by Ball et al. (2003). However, a recent reanalysis of this study (Evans & Ball, 2010) has cast new light on the issue. While matching cards are inspected for much longer amounts of time, they are not always selected, especially in cases where such choices would be hard to justify. This finding accords with other evidence that reasoning does, in fact, affect choices on the task (Feeney & Handley, 2000; Handley, Feeney, & Harper, 2002).

The other main origin of the dual-process theory in the study of deductive reasoning came from the study of syllogistic reasoning and the phenomenon of “belief bias.” The belief bias paradigm provides perhaps the prototypical dual-process finding in the field of deductive reasoning because in contrast with the selection task, it appeared clear from the start that both type 1 and 2 processes were influencing the decisions made. Belief bias is the tendency to judge the validity of a logical argument on the basis of whether one agrees with the conclusion. When strict deductive reasoning instructions are employed, the believability of the conclusion is irrelevant, and so its influence is considered to be a cognitive bias. The effect was first demonstrated by Wilkins (1928), but subsequent research was marred by methodological problems. The modern study of the topic is usually considered to start from the paper of Evans, Barston, and Pollard (1983), who established the basic phenomena with all relevant

controls (replicated many times since; for a recent major study see Klauer, Musch, & Naumer, 2000). The paradigm involves asking people to evaluate syllogisms (three term problems, with two premises and a conclusion) in each of four categories: valid-believable, valid-unbelievable, invalid-believable, and invalid-unbelievable. Thus, conclusions might be supported by logic or by belief or neither. Evans et al. found that participants approved valid conclusions 89% of the time when they were believable but only 56% when unbelievable. Invalid conclusions were accepted only 10% of the time when unbelievable, but an astonishing 71% of the time when they accorded with prior belief. Thus, there were three reliable effects: (a) a preference for valid over invalid conclusions, (b) a preference for believable over unbelievable conclusions, and (c) a belief by logic interaction (more belief bias on invalid arguments). On the basis of protocol analysis and other measures, Evans et al. concluded that there was a *within-participant* conflict between belief and logical reasoning. The same person might go with logic on one problem, but belief on another.

It is interesting to note that dual-process theories arose originally from a more descriptive approach in which both logical and nonlogical factors were observed to influence behavior in these tasks—described as the two-factor theory of reasoning (Evans, 1982). It was only later that they were explicitly linked to “analytic” (type 2) and “heuristic” (type 1) processes (Evans, 1989). Both matching bias and belief bias were seen as nonlogical factors in the two-factor approach. While this chapter is concerned mostly with explanations in terms of two types of cognitive processing, it should be noted that recently there has been some reversion to the two-factor approach, mostly due to a revival in the popularity of mathematical modeling (Beller, Bender, & Kuhnmnnch, 2005; Dube, Rotello, & Heit, 2010; Klauer, Beller, & Hutter, 2010). These recent models treat logical form and prior belief as two *sources* of evidence that influence reasoning responses and have provided convincing support for this claim. A dual-source model is not necessarily a dual-process model, though, as a single type of cognitive process could be influenced by more than one source of evidence. Hence, I will focus here on more cognitively oriented work.

One feature of the protocols collected by Evans et al. (1983) was consistent with those of Wason and Evans (1975): no one mentioned the believability of the conclusion as a reason for his or her decision.

So once again, we have a case where an apparently unconscious type 1 process (belief bias) conflicts with logical reasoning. Even though it is apparently less dominant than matching bias is on the selection task, people once again appear to be unaware of this process and able to rationalize the choices they make. It is on these empirical observations that the dual-process theory of deductive reasoning was formed. Perhaps unsurprisingly, the explanation of cognitive biases exclusively in terms of unconscious or preconscious type 1 processing (Evans, 1989) was the main focus of such research for many years after. Although I now consider that conclusion a mistake that has been corrected in recent writing (Evans, 2006b; 2007a), the case for dual processes made by these early studies remains compelling. It seems that people's choices on these tasks are partly a function of explicit reasoning and partly caused by cognitive biases of which they are unaware.

Evidence for the Reality of the Type 1 and 2 Distinction in the Psychology of Reasoning

Dual-process theories of reasoning, as indicated in the previous section, were developed during the main period of traditional study of deductive reasoning, c. 1960–2000. This was reflected in some of the language originally used, as in the description of logical versus nonlogical factors, or the contrast of the role of logic and belief in syllogistic reasoning. However, it is important to note that not only has dual-process theory adapted to the new paradigm psychology of reasoning, but it has flourished within it (see Evans, *in press*). Dual-process theorists no longer refer to type 2 processing as logical, and the association of type 2 processing with normatively correct solutions has been specifically deemphasized in recent writings (Evans, 2007a; Stanovich, 2010). Nevertheless, the history of these theories has strongly influenced the *received* view of dual-process theory and the persistence of some of the fallacies listed in Table 8.2, which I discuss in the next section.

Dual-process theories are open to criticism on the grounds that the features listed in Table 8.1 are not reliable indicators of two kinds of processing or even consistently correlated with one another. For this reason, recent authors have focussed the debate about the type 1 and 2 distinction on the issue of whether working memory is involved (see Evans, 2003; 2008). By "working memory" I mean a singular, central, and capacity-limited resource of the kind discussed by Baddeley (2007) and as implied

by the large program of work that measures *working memory capacity* and correlates it with many cognitive functions (Barrett, Tugade, & Engle, 2004). Defined in this sense, the contemporary view is that type 2 processing requires and taxes working memory, whereas type 1 processing does not.

As an example, suppose I ask you whether the following statement is believable: "If an animal is a dog, then it has a tail." You might be inclined very quickly to say "yes" without any conscious effort—a rapid, intuitive type 1 response. But if instead I ask you, is it *necessarily* true that "If an animal is a dog, then it has a tail" you might, after reflection, answer no. In the second case you would attempt to call counterexamples to mind: You might think of a breed of dog that has no tail, for example, or consider the possibility that a dog has its tail amputated due to an accident. Reflective thinking of this kind, with consideration of counterexamples (see Johnson-Laird, Chapter 9), requires use of working memory, which makes it a slower and more effortful type 2 process. In fact, recent evidence (Verschueren, Schaeken, & d'Ydewalle, 2005; Oberauer, Geiger, Fischer, & Weidenfeld, 2007) suggests that we may use both associative (type 1) and reflective (type 2) processes when we reason with such conditional statements.

If type 2 processing requires access to working memory, then it explains a number of the features listed in Table 8.1 as being typical of this kind of thought; for example, it is going to be relatively, slow, sequential, high effort and low capacity compared with type 1 thinking, which can bypass working memory. It also explains why type 2 processing is thought to be strongly correlated with measures of cognitive ability. Working memory capacity (WMC) is generally measured by counting the number of items that can be held in short-term memory while simultaneously performing a cognitive task. However, individual differences in WMC are closely correlated with those measured by tests of general intelligence (Colom, Rebollo, Palacios, Juan-Espinosa, & Kyllonen, 2004). Hence, the observation that reasoning ability is closely linked to SAT scores (Stanovich, 1999; Stanovich & West, 2000b), themselves a close correlate of IQ, can readily be interpreted as reflecting the load that type 2 processing puts on working memory.

There are essentially three methods that can be used to test the hypothesis that there are type 2 processes loading working memory, competing or combining with type 1 processes that do not. The

first is to demonstrate that individual differences in cognitive capacity (WMC, IQ, SAT, etc.) *differentially* correlate with behaviors that are attributed to type 1 and 2 processes in standard dual-process accounts of reasoning tasks. The second is to show that performing a reasoning or decision task under a concurrent working memory load *selectively* disrupts type 2 processing. The third method exploits the fact that type 2 processing is generally slower than type 1 processing. So again, we may expect selective disruption of type 2 processes if people are given little time to think—the so-called speeded-task method. Research studies using all three methods appear to support the validity of the type 1 and 2 distinction.

There is a potential problem with the correlational method, however, that should be acknowledged at the outset. The general approach assumes that behaviors dependent upon type 2 processing will be strongly and significantly correlated with cognitive capacity and those based on type 1 processing will have little or no correlation. As an example, Stanovich and West (1998) observed that the small minority of participants who succeed in solving the standard abstract version of the Wason selection task were of exceptionally high IQ. However, when tested on a realistic, deontic version of the task known to be much easier (see Cheng & Holyoak, 1985), there was little correlation with cognitive ability. This was interpreted as evidence that type 2 processing was needed for abstract reasoning, but that the realistic version was facilitated by pragmatic cues of a rapid, automatic (type 1) nature. The problem with the logic here is that a process that uses working memory but does not tax it much may fail to produce a correlation with WMC or one of its correlates in a given sample of people. In fact, Newstead et al. (2004) later showed that with lower ability samples, performance on the deontic selection task *does* correlate substantially with IQ. Even if the correct answer is pragmatically cued on this task, it still requires the ability to follow and comply with a fairly demanding set of instructions, and surely this aspect, at least, loads upon working memory. This raises the issue of whether there are any responses that can be attributed purely and unambiguously to type 1 or 2 processing. Recent research has also complicated the picture in another regard. Whereas earlier studies showed mostly (but not entirely) that individual differences in ability correlated with finding the normatively correct solution to reasoning and

decision problems (Stanovich, 1999), there is now evidence that a substantial range of cognitive biases in decision tasks may be independent of such measures (Stanovich & West, 2008).

The fact that high-ability people can show cognitive biases should not necessarily be seen as evidence against dual-process theory (see discussion of fallacy 3 later). However, it does show how complex the individual differences method is. Stanovich (1999, 2009a, 2009b) has argued that rational performance requires a combination of both sufficient cognitive capacity and a rational thinking disposition. The latter is partly a matter of personality, as measured by scales such as Need for Cognition (Cacioppo & Petty, 1982) but also a function of instructions and context. I will consider two examples, which seem to require retention of a dual-process approach. Stanovich and West (2008) show that some cognitive biases can be resisted by participants of higher cognitive ability provided that a within- rather than between-participant design is adopted. For example, the conjunction fallacy (Tversky & Kahneman, 1983) involves participants judging, in some circumstances, that $P(A \& B)$ is greater than $P(A)$, a logical impossibility. When asked to rate the probabilities in separate groups, cognitive ability is of no help. In a within-participant design, however, high-ability participants may notice the logical inconsistency in their rating of the two outcomes and hence be able to correct it.

Other recent evidence supporting the dual-process approach involves causal conditional inference. This involves study of the four conditional inferences, described earlier, using “if p then q” statements that are realistic and about which people hold prior beliefs of causality. When such materials are used, the endorsement rates for both the valid (MP, MT) and fallacious (DA, AC) inferences may be substantially affected. For example, the tendency to endorse MP can drop well below 100% when people do not believe that p will lead to q in particular conditional statement (e.g. George, 1995; Stevenson & Over, 1995). More generally, when causal conditionals are used, people are substantially influenced by whether p is regarded as a *sufficient* cause for q (affects MP, MT) and as a *necessary* cause for q (affects DA, AC; see Thompson, 1994, 2000). Similar results are found when people are asked to rate the equivalent subjective probabilities: $P(q|p)$ and $P(p|q)$ (e.g., Dieussaert, Schaeken, & d'Ydewalle, 2002). When strict deductive reasoning instructions are employed, as they often are, we can

regard these findings as another form of belief bias. People's prior beliefs about the causal relations are apparently competing with their attempt to follow the instructions and reason deductively.

It has been suggested that the endorsement of conditional inferences is based on nothing more than the subjective conditional probability of the conclusion given the minor premise (Oaksford, Chater, & Larkin, 2000), in which case a dual-process account of causal conditional reasoning would not be required. However, if this were the case, then it should make no difference whether the conditional statement was presented. Consider this DA inference:

If the Bank of England raises interest rates, then the UK inflation rate will decrease.

Suppose the Bank of England does not raise interest rates. Does it follow that the UK inflation rate will not decrease?

If people judge this inference purely on their prior beliefs about causal relations in the economy, then they should give the same answer when the conditional statement is omitted. If, however, they are making an attempt to reason according to instructions, then the fallaciousness of the DA inference should lower their expressed confidence in the conclusion. In a recent study, Klauer et al. (2010) showed that it does make a substantive difference with such inferences whether the conditional statement is present. In a detailed modeling exercise they showed that both the logical form of the inference and prior beliefs influenced the conclusions drawn, in line with a dual-process account.

The dual-process theory is also supported by a recent study of causal conditional inference in my own laboratory (Evans, Handley, Neilens, Bacon, & Over, 2010). We presented such inferences either with pragmatic reasoning instructions (rate the belief in the conclusion given the premises) or with strict deductive reasoning instructions (assume the premises are true and decide whether the conclusion is logically necessary). This is fairly typical of the use of the deduction paradigm within the contemporary psychology of reasoning. It is no longer administered to test people's native logical abilities, but rather as a means of enforcing the need for type 2 processing to deal with a novel and difficult task. In this case, we found that participants of higher IQ were better able to inhibit the influence of belief on their reasoning, but only when deductive reasoning

instructions were given. Under pragmatic instructions, both groups were equally influenced by beliefs. This strongly supports Stanovich's (2009a) argument that both sufficient cognitive ability and a disposition for critical thinking are required for effective type 2 override of type 1 responding. De Neys et al. (2005a, 2005b) similarly found that while higher ability participants are influenced by prior knowledge on such tasks, they will inhibit counterexamples that come to mind which would otherwise block a valid inference. They also showed that reasoning is more belief based when a concurrent working memory load is administered to block type 2 processing.

Turning to the belief bias effect in syllogistic reasoning, there are a number of reported findings consistent with the dual-process theory. Very strong deductive reasoning instructions can reduce but not eliminate the effect (Evans, Allen, Newstead, & Pollard, 1994). When prior belief and logical validity are put into conflict, the ability to inhibit beliefs is greater for those of higher cognitive ability (Stanovich, 1999) but declines sharply in old age (Gilinsky & Judd, 1994). Belief bias is substantially larger when people are forced to respond quickly in a speeded task (Evans & Curtis-Holmes, 2005) or when a concurrent working memory load is administered (De Neys, 2006). These results suggest that belief bias is the result of a rapid type 1 process that provides a default response that may or may not be inhibited and overridden by type 2 reasoning. This conclusion, in turn, is supported by several studies using neural imaging methods. For example, on belief-logic conflict problems, a region in the right prefrontal lateral cortex, associated with inhibitory executive processes, tends to be activated but *only* when the participant gives the answer that favors logic over belief (Goel & Dolan, 2003; Tsujii & Watanabe, 2009).

To summarize, it is difficult to see how a single-process account can be given of the accumulated experimental evidence on deductive reasoning. Even those who advocate a built-in mental logic for reasoning (Braine & O'Brien, 1998a; Rips, 1994) have found the need to augment this with a variety of proposals about nonlogical pragmatic influences, whereas mental model theorists have proposed a mechanism of pragmatic modulation to account for the effects of prior knowledge and belief on deductive inference (Johnson-Laird & Byrne, 2002). It is also clear that attempts to account for behavior in entirely nonlogical terms (Oaksford & Chater,

2007) do not accord with the evidence. The evidence shows that participants do make an effort at deductive reasoning—depending upon their ability and instructions given—and that the logical form of the arguments influences their decisions. In the belief bias paradigm, for example, the validity of the argument appears to compete with the believability of the conclusion, and neural activity differentiates the cases where one or the other wins. On the selection task, people focus on matching cards, but their decisions about them still depend on their logical status (Evans & Ball, 2010). The proposal that type 2 processing is heavily dependent on working memory is supported by converging evidence from studies of individual differences in cognitive capacity, as well as those employing speeded tasks and working memory loads.

Some critics have suggested that type 1 and 2 processing might lie at the ends of a continuum rather than represent distinct types of processing (e.g. Newstead, 2000; Osman, 2004). However, it is hard to see how this can be reconciled with the evidence discussed here. For example, why do patterns of neural activity switch to a different part of the brain when type 2 processing is suggested by the answer given? Although there have been comparatively few imaging studies of reasoning to date (see Goel, 2008 for a recent review), there is extensive evidence from social neuroscience that there are two distinct neural systems underlying implicit and explicit forms of social judgment (Lieberman, 2003, 2007). The pattern of activation and mutual inhibition of these systems is precisely incompatible with the idea that there is just one functional system activated to different degrees (Lieberman, 2009). But even leaving aside the neuroscience, many experimental findings are interactive in nature: cognitive ability, working memory loads, speeded tasks, and instructional manipulations have all been shown *selectively* to affect responses attributed to type 1 and 2 processing.

While the evidence for dual processing remains strong, there are a number of problems with the received view of the theory described at the outset of this chapter. This has led to several fallacies that affect the thinking of both advocates of the theory and its increasing band of critics. It is to these that I now turn. In discussing the fallacies I will refer to not only the core dual-process theory of reasoning whose origins have been traced here but also rather more broadly to dual-process theories in cognitive and social psychology generally.

Fallacies in the Received View of

Dual-Process Theory

Fallacy 1: All Dual-Process Theories Are Essentially the Same

There are many dual-process theories in the cognitive and social psychological literature, but in the *received* view, they tend to be lumped together as incorporating generally the set of features listed in Table 8.1. Some reviewers sympathetic to dual-process theories have sought to integrate a large number of such accounts within a single framework (Evans, 2003; Smith & Collins, 2009; Smith & DeCoster, 2000; Stanovich, 1999). At the same time, critics of dual-process and dual-system theories tend to write as though faults in one account are a problem for all (Gigerenzer & Hoffrage, 2007; Keren & Schul, 2009; Osman, 2004). For example, problems with dual-system theories (see later) can be used as a stick with which to beat all dual-process theories.

As the very least we should distinguish the following: (a) dual-mode theories, (b) dual-type theories, and (c) dual-system theories. In practice, these often get confused. The proposal that there are two *modes* of processing does not necessarily imply that there is more than one kind of cognitive mechanism involved. The proposal of two *types* of thinking does, in the terminology that I am going to use here. It may appear, for example, that the discovery of cross-cultural differences in thinking style (Nisbett, Peng, Choi, & Norenzayan, 2001) must be related to dual-process theories in cognitive and social psychology. Such research indicates that East Asians have a more holistic style of thinking, whereas Westerners are more analytic. But this cannot reflect a type 1 and 2 distinction of the kind I have been discussing in this chapter. First, we can safely assume that all humans share broadly the same cognitive architecture. Second, the research shows that when people move to the other culture, their thinking style accommodates within a matter of months. Attempts to reconcile these findings with dual-process theory include (a) the idea that the tendency to intervene on default intuitions with reflective reasoning may be culturally dependent (Evans, 2009), and (b) the proposal that thinking styles reflect two different kinds of reflective (type 2) thought (Buchel & Norenzayan, 2009).

The same issue arises in social psychology, where authors have distinguished between fast, superficial, and slow reflective ways of processing social information as in the ELM (Petty & Wegener, 1999)

and heuristic-systematic models (Chen & Chaiken, 1999). Some reviewers have treated these theories as similar to type theories (Smith & DeCoster, 2000), whereas others have clearly treated them as mode theories (Evans, 2009; Strack & Deutsch, 2004). There are many ways of engaging in deliberative reasoning, which include being slow and careful, or quick and lazy (see fallacy 5). The type/mode distinction is also clearly reflected in the psychometric approach of Stanovich and West (2000a; Stanovich, 1999). In this work, types of processing are distinguished by cognitive capacity, in the way described earlier, clearly linking type 2 processing to working memory. But the same researchers show that when cognitive ability is factored out, residual variance can be predicted by thinking styles. In more recent writing, Stanovich has emphasized the importance of measuring rational thinking dispositions independently of cognitive ability to get a full picture of someone's capacity for rational decision making (Stanovich, 2009b). Thus, a distinction between people who are applying their cognitive faculties effectively or not is one of modes, not types. And unlike types, modes of processing do indeed seem to lie at two ends of a continuum.

Dual-system theories extend type theories by making specific proposals about the underlying cognitive mechanisms responsible for type 1 and 2 processing. But there are different kinds of dual-system theory as well. Some have a specific domain of reference, such as learning (Sun, Slusarz, & Terry, 2005), reasoning (Sloman, 1996), or social cognition (Lieberman, 2003) to which the authors wish largely to restrict them. The received System 1 and 2 theory, however, implies something that applies much more generally to the mind as a whole (encouraged by such authors as Epstein, 1994; Evans & Over, 1996; Reber, 1993; Stanovich, 1999, 2004). The various authors of the wider theory commit themselves to many similar features, broadly as those listed in Table 8.1. However, it is evident that dual-process theories of particular tasks and paradigms can dispense with many of these added features. If we use working memory involvement as our definition of type 2 processing, for example, we do not need to specify that it is conscious, recently evolved, and distinctively human. Dual-system or two-mind theories (Evans, 2010) are asserted at a higher theoretical level and for a broader purpose: to seek to explain, for example, why dual-process theories are proposed for so many different cognitive tasks. This distinction of level is important. If

critics, for example, find dual-system theories to be vague or unfalsifiable, these features cannot automatically be carried over to more specific and limited dual-process accounts.

Even among dual-process researchers, difference in cognitive architecture between different theories went unnoticed until quite recently. I coined terms "default-interventionist" and "parallel-competitive" to distinguish the two main categories (Evans, 2007b, 2008). Some theories are explicitly parallel, proposing that associative (type 1) and rule-based (type 2) processes operate in parallel, each having a say (Sloman, 1996; Smith & DeCoster, 2000). More commonly, however, dual-process accounts of reasoning and decision making propose that rapid type 1 processes provide a default intuitive response, which may or may not be moderated by subsequent intervention with deliberative, type 2 reasoning (Evans, 2006b; 2007a; Kahneman & Frederick, 2002; Stanovich, 1999, 2009a). It is hard to distinguish these two architectures simply by observing two competing sources of variance (Evans, 2007b), so we need to find indirect means of doing so. It can also be argued that there is room for both kinds of theory in different applications and that the actual architecture of the mind includes different type 1 processes that operate *both* in parallel with and prior to the operation of type 2 thinking (Evans, 2009). However, it is also clear that parallel and sequential models have been applied to the explanation of the same phenomena.

Fallacy 2: There Are Just Two Systems Underlying Type 1 and 2 Reasoning

Dual-process theorists have only themselves to blame for spreading the idea that there are two distinct systems in the mind responsible for type 1 and 2 processing (Epstein, 1994; Evans & Over, 1996; Reber, 1993; Stanovich, 1999), commonly known as System 1 and 2. However, prior to the recent critical attack of Keren and Schul (2009), problems with the two-system idea had already been identified and discussed within the dual-process camp (Evans, 2006a, 2009; Stanovich, 2004, 2009a). The problem lies with the word "system," which suggests a singular and well-defined mechanism. It is more appropriate, in my view, to refer to a two-minds hypothesis, to distinguish the idea that there are forms of cognition which are ancient and shared with other animals, and those that are recently evolved and distinctively human (Evans, 2010). In a two-minds theory, each mind can have

access to multiple systems, in a meaningful sense of that term.

Consider first the idea that there is a System 1 responsible for type 1 processing. There are several kinds of processes that meet the type 1 criteria: associative processes involved in learning and retrieval of implicit knowledge; processes attributed to cognitive modules such as the language acquisition device; preconscious and pragmatic processes that retrieve explicit knowledge for processing through working memory; procedural processes that were once explicitly rehearsed and have become automatic through practice. Is there one system for all of these? Beyond that, is there one system that performs all automatic processing, such as face recognition, memory retrieval, stereotyping, and many others? Of course not. However, dual-process and dual-system theorists have been unspecific about the scope and parameters of System 1 and have at times attributed various different kinds of type 1 process to it. Recognizing this problem, Stanovich (2004) dispensed with the term “System 1” in favor of “TASS”: the set of automated subsystems.

System 2 seems to have a better case for being a system. It is, after all, part of the definition of type 2 processing that it is *singular*, operating in a slow, sequential manner that taxes working memory and allows hypothetical thinking about one possibility at a time (Evans, 2007a). However, is the System 2 that engages in explicit hypothesis testing for Sloman, which intervenes with rational decision making for Kahneman, and which overrides belief bias with abstract reasoning for Evans, just the same system? In fact, is System 2 responsible for all the many cognitive tasks whose performance is correlated with working memory capacity? If so, what exactly is this mighty System 2: working memory, the conscious mind, a homunculus? To assert that either System 2 or working memory is a single system that does reading, reasoning, learning, decision making, and so on is vacuous. So while there is good evidence that there is a type 2 form of cognition, attributing it to a single System 2 does not make a lot of sense. The solution that should be explored, in my view, is the idea that there are multiple *type 2 systems* (Evans, 2009; Samuels, 2009), just as there are multiple type 1 systems. The key difference is that type 2 systems compete for a single central working memory system which they utilize *among other resources*. The evidence suggests that there is not even a single type 2 system for reasoning, as different reasoning tasks recruit a wide variety of brain

regions, according to the exact demands of the task (Goel, 2008).

I think of type 2 systems as ad hoc committees that are put together to deal with a particular problem and then disbanded when the task is completed.¹ Reasoning with abstract and belief-laden syllogisms, for example, recruits different resources, as the neural imaging data indicate: Only the latter involve semantic processing regions of the brain. It is also a fallacy to think of “System 2” as a conscious mind that is choosing its own applications. The ad hoc committee must be put together by some rapid and preconscious process—any feeling that “we” are willing and choosing the course of our thoughts and actions is an illusion (Wegner, 2002). I therefore also take issue with dual-process theorists (e.g., Kahneman & Frederick, 2002; Sloman, 1996) who assign to System 2 not only the capacity for rule-based reasoning but also an overall executive role that allows it to decide whether to intervene upon or overrule a System 1 intuition. In fact, recent evidence suggests that while people’s brains detect conflict in dual-process paradigms, the conscious person does not (De Neys & Glumicic, 2008; De Neys, Vartanian, & Goel, 2008).

Fallacy 3: Type 1 Processes Are Responsible for Cognitive Biases; Type 2 Processes for Normatively Correct Responding

One of the most important fallacies to have arisen in dual-process research is the belief that the normativity of an answer (see Chater & Oaksford, Chapter 2) is diagnostic of the type of processing. Given the history of the dual-process theory of reasoning, one can easily see how this came about. In earlier writing, heuristic or type 1 processes were always the “bad guys,” responsible for cognitive biases (e.g. Evans, 1989). In belief bias research, authors often talked about the conflict between “logic” and “belief,” which are actually dual *sources*, rather than dual processes. Evans and Over (1996) defined “rationality2” as a form of well-justified and explicit rule-based reasoning that could only be achieved by type 2 processes. Stanovich (1999; Stanovich & West, 2000a) in his earlier reviews of his psychometric research program emphasized the association between high cognitive ability, type 2 processing and normative responding. Similarly, Kahneman and Frederick (2002) associate the heuristics of Tversky and Kahneman with System 1 and successful reasoning to achieve normatively correct solutions to the intervention of System 2.

The problem is that a normative system is an externally imposed, philosophical criterion that can have no direct role in the psychological definition of a type 2 process. We may, of course, believe that the brain is adaptive and has evolved in such a way as to make us instrumentally rational, that is, likely to achieve our goals. However, theorists with this approach are sharply divided as to whether a normative theory is required to capture this kind of adaptation (Oaksford & Chater, 2007) or not (Gigerenzer, 2004). Within the psychology of reasoning, the view that there is an in-built mental logic in the form of inference rules (Braine & O'Brien, 1998a; Rips, 1994) is nowadays a minority position, and even these theorists propose substantial variation from standard logics, as well as a host of performance factors. In any case, if type 2 processes are those that manipulate explicit representations through working memory, why should such reasoning necessarily be normatively correct? People may apply the wrong rules or make errors in their application. And why should type 1 processes that operate automatically and without reflection necessarily be wrong? In fact, there is much evidence that expert decision making can often be well served by intuitive rather than reflective thinking (Gigerenzer, 2007; Gladwell, 2005; Klein, 1999; Reyna, 2004) and that sometimes explicit efforts to reason can result in worse performance (Reber, 1993; Wilson & Schooler, 1991).

Reasoning research somewhat loads the dice in favor of type 2 processing by focusing on abstract, novel problems presented to participants without relevant expertise. If a sports fan with much experience of following games is asked to predict results, he or she may be able to do so quite well without need for reflective reasoning. However, a participant in a reasoning experiment is generally asked to do novel things, like assuming some dubious propositions to be true and deciding whether a conclusion necessarily follows from them. In these circumstances, explicit type 2 reasoning is usually necessary for correct solution, but certainly not sufficient. Arguably, however, when prior experience provides appropriate pragmatic cues, even an intractable problem like the Wason selection task becomes easy to solve (Cheng & Holyoak, 1985), as this can be done with type 1 processes (Stanovich & West, 1998). It is when normative performance requires the deliberate suppression of *unhelpful* pragmatic cues that higher ability participants perform better under strict deductive reasoning instructions (De

Neys, Schaeken, & d'Ydewalle, 2005b; Evans et al., 2010).

Hence, fallacy 3 is with us for some fairly precise historical reasons. In the traditional paradigms, researchers presented participants with hard, novel problems for which they lacked experience (students of logic being traditionally excluded), and also with cues that prompted type 1 processes to compete or conflict with these correct answers. So in these paradigms, it does seem that type 2 processing is at least necessary to solve the problems, and that type 1 processes are often responsible for cognitive biases. But this perspective is far too narrow, as has recently been recognized. In recent writing, I have attributed responsibility for a range of cognitive biases roughly equally between type 1 and type 2 processing (Evans, 2007a). Stanovich (2009a, 2010) similarly identifies a number of reasons for error other than a failure to intervene with type 2 reasoning; for example, people may reason in a quick and sloppy (but type 2) manner or lack the necessary "mindware" for successful reasoning.

Fallacy 4: Type 1 Processing Is Contextualized, Whereas Type 2 Processing Is Abstract

Fallacy 4 arises for similar historical reasons to fallacy 3. Belief bias has always featured strongly in dual-process accounts of reasoning, and Stanovich (1999) went so far as to identify contextualization as the fundamental cognitive bias. It does appear that our thinking is automatically contextualized when we are given a reasoning problem which cues prior knowledge, and that participants of higher ability may be better able to inhibit the influence of such context. Interestingly, autistic people are *less* influenced by belief bias in conditional reasoning (McKenzie, Evans, & Hanoch, 2010), probably because their reasoning was not contextualized to start with. In the pre-2000 work on dual-process theory, abstract and logical reasoning became linked with type 2 processing, whereas contextualized and biased reasoning was attributed to type 1 processing, supporting both fallacies 3 and 4.

The signs that something is wrong with this idea have been with us for some time. For example, it has become evident that there is not one but *two* belief bias effects in syllogistic reasoning (Evans, Handley, & Harper, 2001; Klauer et al., 2000; Oakhill, Johnson-Laird, & Garnham, 1989). The first, which we might call a type 1 belief bias (i.e., based on a type 1 process), is a general response bias to say "yes" to conclusions that accord with prior

belief and “no” to those that do not. The type 2 bias affects the process of reasoning such that people search selectively for counterexamples to unbelievable conclusions. These two biases account, respectively, for the main effect and the interaction effect observed in the data. That is, the type 1 bias produces overall preference for believable conclusions, whereas the type 2 bias selectively affects invalid inferences. More recently, there is evidence that beliefs influence causal conditional inference in two similar ways. People are partially influenced by the degree of association between minor premise and conclusion (as suggested by Oaksford et al., 2000) but also by the explicit consideration of counter-examples that come to mind (Verschueren et al., 2005; Weidenfeld, Oberauer, & Hornig, 2005) in line with the general principles of mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991). A recent study suggests that the type 1 belief bias dominates MP and AC inferences, whereas the type 2 effect is stronger for DA and MT (Evans et al., 2010).

A little reflection will reveal why it must be fallacious to equate type 2 processing with abstract, decontextualized thought. Bearing in mind that type 2 processing involves putting a load on working memory—and is not a mental logic—it is evident that much (indeed most) of the contents that pass through our working memories are forms of explicit belief and knowledge that are semantically rich. The dual-process claim must rather be that there are ways in which prior experience can cue intuitions and behavior by associative and procedural learning that bypass working memory (type 1), which are separate from the retrieval and manipulation of explicit memories and beliefs through working memory (type 2). Not only must this be the only coherent claim, but the evidence increasingly supports this view. Fallacy 4 has arisen for the same historical reasons as fallacy 3: When a novel task requires attention to logical structure and hence decontextualization, in general type 2 processing will be necessary (but not sufficient) to achieve this.

Fallacy 5: Fast Processing Indicates Use of a Type 1 Rather Than Type 2 Process

It is widely agreed that type 1 processing is in general quicker than type 2 processing. The speeded-task method exploits this by giving participants a fixed short time limit for a decision or instructing them to respond quickly. In general, this manipulation does seem to shift people toward responses

theoretically assigned to type 1 processing (e.g., belief bias) and has similar effects to use of a concurrent working memory load. Nevertheless, it is a fallacy to assume that because a judgment is made quickly it is necessarily based on a type 1 process.

The term “intuitive judgment” is used rather loosely in the literature, and I suspect it has two different meanings. If intuition means based on feelings without access to explicit reasoning, then that sounds like a type 1 process. But in some applications it seems to mean *naïve* judgement, which could be based on explicit rules or heuristics that occur to an untrained judge, in which case they would be type 2. For example, are the “fast and frugal” heuristics proposed by Gigerenzer and colleagues based on type 1 or 2 processes? Despite the title *Gut Feelings*, Gigerenzer’s (2007) recent book on the topic describes heuristics as simple rules that could well be applied explicitly. (This was clearly the case in the famous application of heuristics in problem solving by Newell & Simon, 1972.) This ambiguity is discussed by Betsch (2008), who suggests a distinction between *intuitive judgment* based on feelings and *heuristic judgment* based on simple rules. The problem is that both may be quick, but in dual-process theory we would have to describe the former as type 1 and the latter as type 2.

Gigerenzer and colleagues frequently use the phrase “less is more” to convey the power of simple heuristics. This perspective is echoed in the popular work *Blink* by Malcolm Gladwell (2005). But consider two of the examples used by Gladwell to illustrate the less-is-more principle: (a) an art expert “knows” that a statue is a fake but cannot prove it by any explicit reasoning or knowledge, and (b) a hospital employs a short checklist to decide whether to treat patients with chest pain as suspected heart attack cases, with much better success than more complex procedures. Both are relatively quick ways of making a decision, but it is evident that judgment (a) has the implicit type 1 characteristics, whereas (b) type 2 is entirely explicit. However, there is an important difference between an intuitive (type 1) and heuristic (type 2) judgement. Both may be less in terms of time taken, but only the latter is less in terms of information considered. As explicit rules, heuristics are by definition quick and dirty, based on minimal cues and requiring little reflection to apply. Intuitive judgements on the other hand may take into account *more* information than could be handled by explicit, type 2 reasoning. This is why it may be harder to learn complex rules explicitly (Reber,

1993) and why expert decision making often relies on intuitive judgement (Klein, 1999; Reyna, 2004). Slow experiential learning may enable a person to take account of *more* cues in the environment, even though these are not consciously accessible.

In summary, type 2 processing is slow when a person reflects on a problem and engages in explicit reasoning or mental simulation to try to solve it. Intuitive (type 1) judgments will generally be quick, even though they may be based on a large amount of information if it is of the kind that can be encoded by associative learning. The problem is that there are also fast type 2 judgements that are made on the basis of simple rules and heuristics, with minimal reflection. This kind of type 2 processing will also make minimal demands on working memory and so is unlikely to reveal itself in a correlation with measures of cognitive ability or capacity. However, when people are given novel tasks—for which they are unlikely to have any available explicit heuristics—requiring them to respond quickly is still likely to shift the balance of power in favor of type 1 processing, as the research evidence suggests.

Conclusions

From around 1960 to 2000 the psychology of reasoning was strongly focused on the deduction paradigm, in which participants are assessed for their ability to judge the logical validity of arguments without any prior training or instruction. In the 1980s a major argument was developed about whether deductive competence was based on built-in inference rules (mental logic) or on the manipulation of mental models. However, experimental evidence has accumulated indicating that people were apparently poor at logical reasoning, highly influenced by irrelevant features of the content and subject to a number of other cognitive biases. As a result, the paradigm has shifted in the past 10–15 years, and while deduction is still studied, it is allied to a number of other methods. Theoretically, there is now also a lot more interest in human reasoning as a probabilistic and pragmatic, rather than deductive process (see Hahn & Oaksford, Chapter 15).

Dual-process theories are the focus of much contemporary research in cognitive and social psychology and have a number of (largely independent) origins. In this chapter, I have shown how dual-process accounts were developed from the 1970s onward within the study of deductive reasoning, primarily as an attempt to explain why cognitive

biases coexisted and competed with attempts at effortful logical reasoning on these tasks. In more recent times, reasoning researchers have examined individual differences in cognitive capacity, linked with type 2 (analytic) reasoning, and also introduced a range of experimental and neuroscientific methods to identify dual processes. However, I have also shown how a number of fallacies have arisen in the received view of dual-process theory, both within the psychology of reasoning and more generally. Five of the more important ones (Table 8.2) have been discussed in detail.

Future Directions

Research on dual-process theory remains a hot topic in cognitive psychology, and it is encouraging that linkage between such theories in reasoning, decision making, learning, and social cognition is now being recognized by contemporary authors. However, dual-process theory is also controversial and coming under increasing criticism from those unsympathetic to the approach. While I believe that the evidence for dual process is strong, there is also need for researchers to be a good deal more careful in formulating proposals and interpreting evidence in this field. The *received* view of the theory, which emerged around 2000 and is summarized in Table 8.1, is fraught with difficulties and has spawned a number of fallacies, which I have discussed here. Dual-process theories should not be (but often are) confused with accounts which simply propose two modes of processing that might be operated by a single cognitive mechanism, or two cognitive styles that may be culturally relative. Theories that genuinely posit two types of processing may also differ in cognitive architecture and in whether they subscribe to broader, two-system proposals.

To minimize the problems posed by the fallacies, I suggest the only firm foundation for two types of processing is that one can bypass working memory, whereas the other necessarily requires its use. However, the fallacies discussed here nevertheless reveal profound methodological difficulties for the study of dual processes in human reasoning. We cannot simply diagnose a type 2 process when we see a normatively correct answer, or a type 1 process when we see a bias. A process that is fast and fails to tax working memory sufficiently to correlate with cognitive ability may, nevertheless, be type 2 in some cases. Similarly, not all answers influenced or biased by belief are the result of type 1 processes.

The difficulties are so great that the reader may wonder why I continue to advocate a dual-process account at all. The earlier sections of this chapter make the case, however. First, the evidence for dual sources of variance in reasoning tasks is beyond dispute. It was the competing influence of nonlogical and task-irrelevant sources, resulting in matching and belief bias, which first pointed to the need for a dual-process account, and it still does. Recent mathematical modeling of these tasks confirms dual sources by several different methods. Problematic as they are in some respects, the standard methods involving working memory loads, correlation with cognitive capacity, and speeded tasks generate findings that are remarkably consistent with original dual-process claims when applied to these paradigms and others. There is strong behavioral and neuroscientific evidence that the brain has parallel implicit and explicit learning systems, and a clear evolutionary argument for the two-minds hypothesis (Evans, 2010; Stanovich, 2004). Crucially, perhaps, neuroimaging studies have confirmed that responding attributed to type 1 and 2 processes corresponds to activation of quite distinct regions in the brain, as well as monitoring both the detection and resolution of conflict.

Accordingly, my conclusion is that one-process theories cannot account for the massive amount of evidence now collected by dual-process researchers. Although the terms “System 1” and “System 2” now seem unhelpful, the case for a two-minds theory incorporating the distinction between type 1 and type 2 systems is a good deal more coherent. However, the case for dual-process theories is not helpfully advanced by researchers who implicitly or explicitly adopt any of the five fallacies that I have discussed in this chapter. Greater care needs to be taken in both the design of research studies and in the interpretation of their findings. Where possible, conclusions reached on the basis of one behavioral measure should be cross-validated with others and, if possible, by neuroimaging studies. This is a challenge to which I hope the psychology of reasoning will be able to rise in the years ahead.

Acknowledgments

I would like thank Keith Frankish and Shira Elqayam for their critical reading of an earlier draft of this chapter.

Note

1. Some philosophers have suggested that “System 2” might be an emergent property of the interactions of multiple modules or subsystems (Carruthers, 2006; Frankish, 2004), which sounds

a similar idea to ad hoc committees. However, my proposal differs first in regarding working memory as a real, not a virtual system, which is an essential component of all type 2 systems, and second by the proposal of unique recruitment of the particular resources required for the task at hand.

References

- Baddeley, A. (2007). *Working memory, thought and action*. Oxford, England: Oxford University Press.
- Ball, L. J., Lucas, E. J., Miles, J. N. V., & Gale, A. G. (2003). Inspection times and the selection task: What do eye-movements reveal about relevance effects? *Quarterly Journal of Experimental Psychology*, 56A, 1053–1077.
- Barrett, L. F., Tugade, M. M., & Engle, R. W. (2004). Individual differences in working memory capacity and dual-process theories of the mind. *Psychological Bulletin*, 130, 553–573.
- Beller, S., Bender, A., & Kuhmnnch, G. (2005). Understanding conditional promises and threats. *Thinking and Reasoning*, 11, 209–238.
- Betsch, T. (2008). The nature of intuition and its neglect in research on judgement and decision making. In H. Plessner, C. Betsch, & T. Betsch (Eds.), *Intuition in judgment and decision making* (pp. 3–22). New York: Erlbaum.
- Braine, M. D. S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review*, 85, 1–21.
- Braine, M. D. S., & O’Brien, D. P. (Eds.). (1998a). *Mental logic*. Mahwah, NJ: Erlbaum.
- Braine, M. D. S., & O’Brien, D. P. (1998b). The theory of mental-propositional logic: Description and illustration. In M. D. S. Braine & D. P. O’Brien (Eds.), *Mental logic* (pp. 79–89). Mahwah, NJ: Erlbaum.
- Buchel, C., & Norenzayan, A. (2009). Thinking across cultures: Implications for dual processes. In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 217–238). Oxford, England: Oxford University Press.
- Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, 42, 116–131.
- Carruthers, P. (2006). *The architecture of the mind*. Oxford, England: Oxford University Press.
- Chen, S., & Chaiken, S. (1999). The heuristic-systematic model in its broader context. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 73–96). New York: The Guilford Press.
- Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, 17, 391–416.
- Colom, R., Rebollo, I., Palacios, A., Juan-Espinosa, M., & Kyllonen, P. C. (2004). Working memory is (almost) perfectly predicted by g. *Intelligence*, 32, 277–296.
- Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? *Cognition*, 31, 187–276.
- De Neys, W. (2006). Dual processing in reasoning - Two systems but one reasoner. *Psychological Science*, 17, 428–433.
- De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, 106, 1248–1299.
- De Neys, W., Schaeken, W., & d’Ydewalle, G. (2005a). Working memory and counterexample retrieval for causal conditionals. *Thinking and Reasoning*, 11, 123–150.
- De Neys, W., Schaeken, W., & d’Ydewalle, G. (2005b). Working memory and everyday conditional reasoning: Retrieval and inhibition of stored counterexamples. *Thinking and Reasoning*, 11, 349–381.

- De Neys, W., Vartanian, O., & Goel, V. (2008). Smarter than we think: When our brains detect that we are biased. *Psychological Science*, 19, 483–489.
- Deutsch, R., & Strack, F. (2006). Reflective and impulsive determinants of addictive behavior. In W. W. Reinout & A. W. Stacy (Eds.), *Handbook of implicit cognition and addiction* (pp. 45–57). Thousand Oaks, CA.: Sage.
- Dieussaert, K., Schaeken, W., & d'Ydewalle, G. (2002). The relative contribution of content and context factors on the interpretation of conditionals. *Experimental Psychology*, 49, 181–195.
- Dube, C., Rotello, C., & Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review*, xx, xx.
- Epstein, S. (1994). Integration of the cognitive and psychodynamic unconscious. *American Psychologist*, 49, 709–724.
- Evans, J. St. B. T. (1982). *The psychology of deductive reasoning*. London: Routledge.
- Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences*. Brighton, England: Erlbaum.
- Evans, J. St. B. T. (1996). Deciding before you think: Relevance and reasoning in the selection task. *British Journal of Psychology*, 87, 223–240.
- Evans, J. St. B. T. (1998). Matching bias in conditional reasoning: Do we understand it after 25 years? *Thinking and Reasoning*, 4, 45–82.
- Evans, J. St. B. T. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin*, 128, 978–996.
- Evans, J. St. B. T. (2003). In two minds: Dual process accounts of reasoning. *Trends in Cognitive Sciences*, 7, 454–459.
- Evans, J. St. B. T. (2006a). Dual system theories of cognition: Some issues. In *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*. Vancouver. Retrieved August 2011, from: <http://www.cogsci.rpi.edu/CSJarchive/proceedings/2006/docs/p202.pdf>
- Evans, J. St. B. T. (2006b). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin and Review*, 13, 378–395.
- Evans, J. St. B. T. (2007a). *Hypothetical thinking: Dual processes in reasoning and judgement*. Hove, England: Psychology Press.
- Evans, J. St. B. T. (2007b). On the resolution of conflict in dual-process theories of reasoning. *Thinking and Reasoning*, 13, 321–329.
- Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment and social cognition. *Annual Review of Psychology*, 59, 255–278.
- Evans, J. St. B. T. (2009). How many dual-process theories do we need: One, two or many? In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 31–54). Oxford, England: Oxford University Press.
- Evans, J. St. B. T. (in press). Reasoning. In D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology* (pp. xx). New York: Oxford University Press.
- Evans, J. St. B. T., Allen, J. L., Newstead, S. E., & Pollard, P. (1994). Debiasing by instruction: The case of belief bias. *European Journal of Cognitive Psychology*, 6, 263–285.
- Evans, J. St. B. T., & Ball, L. J. (2010). Do people reason on the Wason selection task: A new look at the data of Ball et al. (2003). *Quarterly Journal of Experimental Psychology*, 63, 434–441.
- Evans, J. St. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory and Cognition*, 11, 295–306.
- Evans, J. St. B. T., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking and Reasoning*, 11, 382–389.
- Evans, J. St. B. T., Handley, S. J., & Harper, C. (2001). Necessity, possibility and belief: A study of syllogistic reasoning. *Quarterly Journal of Experimental Psychology*, 54A, 935–958.
- Evans, J. St. B. T., Handley, S., Neilens, H., Bacon, A. M., & Over, D. E. (2010). The influence of cognitive ability and instructional set on causal conditional inference. *Quarterly Journal of Experimental Psychology*, 63, 892–909.
- Evans, J. St. B. T., & Lynch, J. S. (1973). Matching bias in the selection task. *British Journal of Psychology*, 64, 391–397.
- Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction*. Hove, England: Erlbaum.
- Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, England: Psychology Press.
- Evans, J. St. B. T., & Over, D. E. (2004). *If*. Oxford, England: Oxford University Press.
- Evans, J. St. B. T., & Wason, P. C. (1976). Rationalisation in a reasoning task. *British Journal of Psychology*, 63, 205–212.
- Feeney, A., & Handley, S. J. (2000). The suppression of q card selections: Evidence for deductive inference in Wason's selection task. *Quarterly Journal of Experimental Psychology*, 53A, 1224–1243.
- Fiddick, L., Cosmides, L., & Tooby, J. (2000). No interpretation without representation: The role of domain-specific representations and inferences in the Wason selection task. *Cognition*, 77, 1–79.
- Frankish, K. (2004). *Mind and supermind*. Cambridge, England: Cambridge University Press.
- Frankish, K., & Evans, J. St. B. T. (2009). The duality of mind: An historical perspective. In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 1–30). Oxford, England: Oxford University Press.
- George, C. (1995). The endorsement of the premises: Assumption-based or belief-based reasoning. *British Journal of Psychology*, 86, 93–111.
- Gigerenzer, G. (2004). Fast and frugal heuristics: The tools of bounded rationality. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 62–88). Oxford, England: Blackwell.
- Gigerenzer, G. (2007). *Gut feelings*. London: Penguin.
- Gigerenzer, G., & Hoffrage, U. (2007). The role of representation in Bayesian reasoning: Correcting common misconceptions. *Behavioral and Brain Sciences*, 30, 264–267.
- Gilinsky, A. S., & Judd, B. B. (1994). Working memory and bias in reasoning across the life-span. *Psychology and Aging*, 9, 356–371.
- Gladwell, M. (2005). *Blink*. London: Penguin.
- Goel, V. (2008). Anatomy of deductive reasoning. *Trends in Cognitive Sciences*, 11, 435–441.
- Goel, V., & Dolan, R. J. (2003). Explaining modulation of reasoning by belief. *Cognition*, 87, B11–B22.
- Handley, S. J., Feeney, A., & Harper, C. (2002). Alternative antecedents, probabilities and the suppression of fallacies on Wason's selection task. *Quarterly Journal of Experimental Psychology*, 55A, 799–813.
- Henle, M. (1962). On the relation between logic and thinking. *Psychological Review*, 69, 366–378.

- Holyoak, K., & Cheng, P. (1995). Pragmatic reasoning with a point of view. *Thinking and Reasoning*, 1, 289–314.
- Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking*. New York: Basic Books.
- Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, England: Cambridge University Press.
- Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hove & London: Erlbaum.
- Johnson-Laird, P. N., & Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics and inference. *Psychological Review*, 109, 646–678.
- Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgement. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge, England: Cambridge University Press.
- Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on psychological science*, 4, 533–550.
- Klauer, K. C., Beller, S., & Hutter, M. (2010). Conditional reasoning in context: A dual-source model of probabilistic inference. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 36, 298–323.
- Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review*, 107, 852–884.
- Klein, G. (1999). *Sources of power*. Cambridge, MA: MIT Press.
- Lieberman, M. D. (2003). Reflective and reflexive judgment processes: A social cognitive neuroscience approach. In J. P. Forgas, K. R. Williams, & W. von Hippel (Eds.), *Social judgments: Implicit and explicit processes* (pp. 44–67). New York: Cambridge University Press.
- Lieberman, M. D. (2007). Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology*, 58, 259–289.
- Lieberman, M. D. (2009). What zombies can't do: A social cognitive neuroscience approach to the irreducibility of reflective consciousness. In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 293–316). Oxford, England: Oxford University Press.
- Lucas, E. J., & Ball, L. J. (2005). Think-aloud protocols and the selection task: Evidence for relevance effects and rationalisation processes. *Thinking and Reasoning*, 11, 35–66.
- Manktelow, K. I. (1999). *Reasoning and thinking*. Hove, England: Psychology Press.
- Manktelow, K. I., & Over, D. E. (1991). Social roles and utilities in reasoning with deontic conditionals. *Cognition*, 39, 85–105.
- McKenzie, R., Evans, J. St. B. T., & Hanoch, Y. (2010). Conditional reasoning in autism: Activation and integration of knowledge and belief. *Developmental Psychology*, 46, 391–403.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
- Newstead, S. E. (2000). Are there two different kinds of thinking? *Behavioral and Brain Sciences*, 23, 690–691.
- Newstead, S. E., Handley, S. J., Harley, C., Wright, H., & Farely, D. (2004). Individual differences in deductive reasoning. *Quarterly Journal of Experimental Psychology*, 57A, 33–60.
- Nisbett, R., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic vs analytic cognition. *Psychological Review*, 108, 291–310.
- Oakhill, J., Johnson-Laird, P. N., & Garnham, A. (1989). Believability and syllogistic reasoning. *Cognition*, 31, 117–140.
- Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101, 608–631.
- Oaksford, M., & Chater, N. (2007). *Bayesian rationality*. Oxford, England: Oxford University Press.
- Oaksford, M., & Chater, N. (Eds.). (2010). *Cognition and conditionals*. Oxford, England: Oxford University Press.
- Oaksford, M., Chater, N., & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 26, 883–889.
- Oberauer, K., Geiger, S. M., Fischer, K., & Weidenfeld, A. (2007). Two meanings of 'If': Individual differences in the interpretation of conditionals. *Quarterly Journal of Experimental Psychology*, 60, 790–819.
- Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin and Review*, 11, 988–1010.
- Over, D. E., Evans, J. St. B. T., & Elqayam, S. (2010). Conditionals and non-constructive reasoning. In M. Oaksford & N. Chater (Eds.), *Cognition and conditionals: Probability and logic in human thinking* (pp. 131–151). Oxford, England: Oxford University Press.
- Petty, R. E., & Wegener, D. T. (1999). The elaboration likelihood model: Current status and controversies. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 41–72). New York: The Guilford Press.
- Reber, A. S. (1993). *Implicit learning and tacit knowledge*. Oxford, England: Oxford University Press.
- Reyna, V. F. (2004). How people make decisions that involve risk: A dual-processes approach. *Current Directions in Psychological Science*, 13, 60–66.
- Rips, L. J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.
- Roberts, M. J. (1998). Inspection times and the selection task: Are they relevant? *Quarterly Journal of Experimental Psychology*, 51A, 781–810.
- Samuels, R. (2009). The magic number two plus or minus: Some comments on dual-processing theories of cognition. In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 129–146). Oxford, England: Oxford University Press.
- Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119, 3–22.
- Smith, E. R., & Collins, E. C. (2009). Dual-process models: A social psychological perspective. In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 197–216). Oxford, England: Oxford University Press.
- Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review*, 4, 108–131.
- Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition*, 57, 31–95.
- Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Stanovich, K. E. (2004). *The robot's rebellion: Finding meaning in the age of Darwin*. Chicago, IL: Chicago University Press.
- Stanovich, K. E. (2009a). Distinguishing the reflective, algorithmic and autonomous minds: Is it time for a tri-process theory? In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 55–88). Oxford, England: Oxford University Press.
- Stanovich, K. E. (2009b). *What intelligence tests miss. The psychology of rational thought*. New Haven, CT & London: Yale University Press.

- Stanovich, K. E. (2010). *Rationality and the reflective mind*. New York: Oxford University Press.
- Stanovich, K. E., & West, R. F. (1998). Cognitive ability and variation in selection task performance. *Thinking and Reasoning*, 4, 193–230.
- Stanovich, K. E., & West, R. F. (2000a). Advancing the rationality debate. *Behavioral and Brain Sciences*, 23, 701–726.
- Stanovich, K. E., & West, R. F. (2000b). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences*, 23, 645–726.
- Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, 94, 672–695.
- Stevenson, R. J., & Over, D. E. (1995). Deduction from uncertain premises. *Quarterly Journal of Experimental Psychology*, 48A, 613–643.
- Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, 8, 220–247.
- Sun, R., Slusarz, P., & Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: A dual-process approach. *Psychological Review*, 112, 159–192.
- Thompson, V. A. (1994). Interpretational factors in conditional reasoning. *Memory and Cognition*, 22, 742–758.
- Thompson, V. A. (2000). The task-specific nature of domain-general reasoning. *Cognition*, 76, 209–268.
- Tsuji, T., & Watanabe, S. (2009). Neural correlates of dual-task effect on belief-bias syllogistic reasoning: A near-infrared spectroscopy study. *Brain Research*, 1287, 118–125.
- Tversky, A., & Kahneman, D. (1983). Extensional vs intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90, 293–315.
- Verschueren, N., Schaeken, W., & d'Ydewalle, G. (2005). A dual-process specification of causal conditional reasoning. *Thinking and Reasoning*, 11, 239–278.
- Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology I* (pp. 106–137). Harmondsworth, England: Penguin.
- Wason, P. C., & Evans, J. St. B. T. (1975). Dual processes in reasoning? *Cognition*, 3, 141–154.
- Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content*. London: Batsford.
- Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, MA: MIT books.
- Weidenfeld, A., Oberauer, K., & Hornig, R. (2005). Causal and noncausal conditionals: An integrated model of interpretation and reasoning. *Quarterly Journal of Experimental Psychology*, 58, 1479–1513.
- Wilkins, M. C. (1928). The effect of changed material on the ability to do formal syllogistic reasoning. *Archives of Psychology*, 16, 142.
- Wilson, T. D. (2002). *Strangers to ourselves*. Cambridge, MA: Belknap Press.
- Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, 60, 181–192.
- Woodworth, R. S., & Sells, S. B. (1935). An atmosphere effect in syllogistic reasoning. *Journal of Experimental Psychology*, 18, 451–460.

Inference with Mental Models

P. N. Johnson-Laird

Abstract

This chapter outlines the theory of mental models. The theory accounts for the deductive reasoning of individuals untrained in logic, and the chapter marshals corroboratory evidence. It goes on to describe how the theory applies to reasoning about probabilities based on the alternative possibilities in which an event occurs (“extensional” reasoning). It also outlines the theory’s application to inductions, including those that depend on the intuitive system of reasoning (System 1) and those that depend on deliberation (System 2). Inductions include the automatic use of knowledge to modulate the interpretation of assertions. Models appear to underlie both the detection of inconsistencies among propositions and abductions that create explanations, including those that resolve inconsistencies. The model theory is supposed to apply to all thinking about propositions, and the chapter concludes with some of the main gaps in its account.

Key Words: abduction, causation, deduction, induction, inconsistency, mental models, reasoning

Introduction

How do we think? One answer is that we rely on *mental models*. Perception yields models of the world that lies outside us. An understanding of discourse yields models of the world that the speaker describes to us. And thinking, which enables us to anticipate the world and to choose a course of action, relies on internal manipulations of these mental models. The present chapter is about this theory, which it refers to as the *model theory*, and about its experimental corroborations. The theory aims to explain all sorts of thinking about propositions, that is, thoughts capable of being true or false. There are other sorts of thinking—the thinking, for instance, of a musician who is improvising. In daily life, unlike the psychological laboratory, no clear demarcation exists between one sort of thinking and another. Here is a sequence of everyday thoughts that illustrates this point:

I had the book in the hotel’s restaurant, and now I’ve lost it. So either I left it in the restaurant, or it fell

out of my pocket on the way back to my room, or it’s somewhere here in my room. It couldn’t have fallen from my pocket—my pockets are deep and I walked slowly back to my room—and so it’s here or in the restaurant.

Embedded in this sequence is a logical deduction of this sort:

A or B or C.

Not B.

Therefore, A or C.

The conclusion is *valid*: It must be true given that the premises are true. But other sorts of thinking occur in the protocol, for example, the inference that the book couldn’t have fallen out of the protagonist’s pocket.

A simple way to categorize thinking about propositions is in terms of its effects on the possibilities consistent with the premises and with the conclusions (Johnson-Laird, 2006). The more possibilities

an assertion rules out, the greater the amount of semantic information it conveys (Bar-Hillel & Carnap, 1964). Any step in thought from current premises to a new conclusion therefore falls into one of five cases:

- The premises and the conclusion are consistent with the same possibilities.
- The premises hold in all the possibilities consistent with the conclusion, but it is consistent with additional possibilities in which the premises do not hold.
- The conclusion holds in all the possibilities consistent with the premises, but they hold in additional possibilities in which the conclusion does not hold.
- The premises and conclusion hold in overlapping possibilities.
- The premises and conclusion hold in completely disjoint possibilities.

The first two cases are valid deductions: The conclusion is true given the truth of premises (see Evans, Chapter 8). The third case includes the traditional instances of induction: The conclusion goes beyond the information given to exclude possibilities otherwise consistent with the premises (see the chapters in Part II of this volume). The fourth case occurs when the conclusion is consistent with the premises, but it does not follow from them. Such thinking occurs in an abduction that explains the premises at the cost of refuting one or more of them (see Lombrozo, Chapter 14). The fifth case occurs when the premises and conclusion are inconsistent.

The model theory aims to explain thinking based on propositions (see Markman, Chapter 4), and the present chapter illustrates its application to the five preceding categories. It begins with the history of the model theory. It then outlines the current theory and its account of deduction. It reviews the evidence for this account. It shows how the theory extends to probabilistic reasoning. It then turns to induction, including the unconscious inductions that affect the interpretation of assertions. It examines how individuals detect inconsistencies and how they make abductions, such as those that create explanations resolving inconsistencies. Finally, it assesses the current state of the model theory and the principal tasks for future research.

The History of Mental Models

In the fifth chapter of his book *The Nature of Explanation*, Kenneth Craik (1943) wrote:

If the organism carries a “small-scale model” of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and the future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it. (p. 61)

As Craik wrote, this same process of internal imitation of the external world is carried out in mechanical devices such as Kelvin's tidal predictor. Craik died in 1945 before he could develop his ideas, and he believed that reasoning depended on verbal rules rather than on models. Several earlier thinkers had, in fact, anticipated the idea of models (see Johnson-Laird, 2004). Nineteenth-century physicists, including Kelvin, Boltzmann, and Maxwell, stressed the role of models in thinking. In the 20th century, physicists downplayed these ideas with the advent of quantum theory (but cf. Deutsch, 1997).

One principle of the modern theory is that the parts of a mental model and their structural inter-relations correspond to those of the situation that it represents. This idea has many antecedents. It occurs in Maxwell's (1911) views on diagrams, in Wittgenstein's (1922) “picture” theory of meaning, and in Köhler's (1938) hypothesis of an isomorphism between brain fields and the world. However, the 19th-century grandfather of the model theory is Charles Sanders Peirce. He co-invented the main system of logic known as the *predicate calculus*, which governs sentences in a formal language containing idealized versions of negation, sentential connectives such as “and” and “or,” and quantifiers such as “all” and “some” (see Chater & Oaksford, Chapter 2). Peirce devised two diagrammatic systems of reasoning, not to improve reasoning, but to display its underlying mental steps (see Johnson-Laird, 2002). He wrote:

Deduction is that mode of reasoning which examines the state of things asserted in the premisses, forms a diagram of that state of things, perceives in the parts of the diagram relations not explicitly mentioned in the premisses, satisfies itself by mental experiments upon the diagram that these relations would always subsist, or at least would do so in a certain proportion of cases, and concludes their necessary, or probable, truth. (Peirce, 1.66; this standard notation refers to paragraph 66 of Volume 1 of Peirce, 1931–1958)

Diagrams can be *iconic*, that is, have the same structure as what they represent (Peirce, 4.447). It is the inspection of an iconic diagram that reveals truths other than those of the premises (2.279, 4.530). Hence, Peirce anticipates Maxwell, Wittgenstein, Köhler, and the model theory. Mental models are as iconic as possible (Johnson-Laird, 1983; see also Markman, Chapter 4).

A resurgence of mental models in cognitive science began in the 1970s. Theorists proposed that knowledge was represented in mental models, but they were not wedded to any particular structure for models. Hayes (1979) used the predicate calculus to describe the naive physics of liquids. Other theorists in artificial intelligence proposed accounts of how to envision models and to use them to simulate behavior (de Kleer, 1977). Psychologists similarly examined naive and expert models of various domains, such as mechanics (McCloskey, Caramazza, & Green, 1980) and electricity (Gentner & Gentner, 1983). They argued that vision yields a mental model of the three-dimensional structure of the world (Marr, 1982). They proposed that individuals use these models to simulate behavior (Hegarty, 1992; Schwartz & Black, 1996; see Hegarty & Stull, Chapter 31). They also studied how models develop (e.g., Halford, 1993; Vosniadou & Brewer, 1992), how they serve as analogies (e.g., Holland, Holyoak, Nisbett, & Thagard, 1986; see Holyoak, Chapter 13), and how they help in the diagnosis of faults (e.g., Rouse & Hunt, 1984). Artifacts, they argued, should be designed so that users easily acquire models of them (e.g., Ehrlich, 1996; Moray, 1990, 1999).

Discourse enables humans to experience the world by proxy, and so an early hypothesis was that comprehension yields models of the world (Johnson-Laird, 1970). Such models are iconic in these ways: They contain a token for each referent in the discourse, properties of these tokens correspond to the properties of the referents, and relations among the tokens correspond to the relations among the referents. Similar ideas occurred in psycholinguistics (e.g., Bransford, Barclay, & Franks, 1972), linguistics (Karttunen, 1976), artificial intelligence (Webber, 1978), and formal semantics (Kamp, 1981). Experimental evidence corroborated the hypothesis, showing that individuals rapidly forget surface and underlying syntax (Johnson-Laird & Stevenson, 1970), and even the meaning of individual sentences (Garnham, 1987). They retain models of only such matters as who did what to whom.

Psycholinguists showed that models are constructed from the meanings of sentences, general knowledge, and knowledge of human communication (e.g., Garnham, 2001; Garnham & Oakhill, 1996; Gernsbacher, 1990; Glenberg, Meyer, & Lindem, 1987). Another early discovery was that content affects deductive reasoning (Wason & Johnson-Laird, 1972), which was hard to reconcile with the then dominant view that reasoners depend on *formal* rules of inference (Braine, 1978; Johnson-Laird, 1975; Osherson, 1974–1976). Granted that models come from perception and discourse, they could be used to reason (Johnson-Laird, 1975): An inference is *valid* if its conclusion holds in all the models of the premises, because its conclusion must be true granted that its premises are true. The next section spells out this theory.

Models and Deduction

Mental models represent entities and persons, events and processes, and the operations of complex systems. But what is a mental model? The current theory is based on three principles that distinguish models from linguistic structures, semantic networks, and other proposed mental representations (cf. Markman, Chapter 4). The first principle (Johnson-Laird, 1983) is as follows:

- I. The principle of *iconicity*: A mental model has a structure that corresponds to the known structure of what it represents.

Visual images are iconic, but mental models underlie images (see Hegarty & Stull, Chapter 31). Even the rotation of mental images implies that individuals rotate three-dimensional models (Metzler & Shepard, 1982). Moreover, models of abstract entities cannot be visualized even though the models may be iconic. To illustrate this point, consider the three-dimensional representation of the world that congenitally blind individuals construct. As Landau, Spelke, and Gleitman (1984) showed, a blind child inferred new routes between objects as a result of walking between them. Her model of the world was iconic within a framework of spatial coordinates, but not visual. Similarly, an iconic model can represent that one set of entities overlaps with another set, and it can do so without any concomitant visual image. Indeed, evidence shows that when individuals reason from materials that do elicit visual images, their reasoning is slower than from other sorts of contents (Knauff, Fangmeier, Ruff, & Johnson-Laird, 2003; Knauff & Johnson-Laird, 2002).

One advantage of iconicity, as Peirce noted, is that models built from premises can yield new relations. For example, Schaeken, Johnson-Laird, and d'Ydewalle (1996) investigated problems of temporal reasoning concerning such propositions as:

John eats his breakfast before he listens to the radio.

Given a problem based on several premises of this sort:

- A before B.
- B before C.
- D while A.
- E while C.

Reasoners can build a mental model with the iconic structure:

A	B	C
D		E

where the left-to-right axis is time, and the vertical axis allows different events to be contemporaneous. Granted that each event takes roughly the same amount of time, reasoners can infer a new relation:

D before E.

Formal logic less readily yields the conclusion, because infinitely many conclusions follow validly from any set of premises, and logic does not tell you *which* conclusions are useful. From the aforementioned premises, for instance, this redundant conclusion follows:

A before B, *and* B before C.

Possibilities are crucial in human thinking, and the second principle of the theory assigns them a central role (Johnson-Laird & Byrne, 1991):

II. The principle of *possibilities*: Each mental model represents a distinct possibility, that is, it captures what is common to all the different ways in which the possibility might occur.

This principle is illustrated in *sentential* reasoning, which hinges on negation and such sentential connectives as "if" and "or." In logic, these connectives have idealized meanings: They are *truth-functional* in that the truth-values of sentences formed with them depend solely on the truth-values of the clauses that they connect. For example, a disjunction of the form: *A or else B but not both*, is true if *A* is true and *B* is false, and if *A* is false and *B* is true, but false in any

other case. This disjunction is known as "exclusive," because it excludes the truth of both *A* and *B*. Logicians capture these conditions in a truth table, as shown in Table 9.1. Each row in the table represents a different possibility; for example, the first row represents the possibility in which both *A* and *B* are true, and so in this case the disjunction is false.

Naive reasoners do not use truth tables (Osherson, 1974–1976). *Fully explicit* models of possibilities, however, are a step toward psychological plausibility. The fully explicit models of the exclusive disjunction, *There is a king or else there is an ace, but not both*, are shown here on two separate horizontal lines:

ace	not-king
not-ace	king

where "not" denotes negation. Table 9.2 presents the fully explicit models of assertions based on the main sentential connectives. Fully explicit models correspond exactly to the true rows in the truth table for each connective. As Table 9.2 shows, the conditional *If A then B* is treated in logic as though it can be paraphrased as *If A then B, and if not-A then B or not-B*. The paraphrase does not do justice to the varied meanings of everyday conditionals (Johnson-Laird & Byrne, 2002). In fact, as we will see, no connectives in natural language can be captured in truth tables.

Fully explicit models yield a more efficient reasoning procedure than truth tables. Each premise has a set of fully explicit models, for example, the premises:

1. There is a king or else there is an ace but not both.
2. There is not a king.

have the models:

Table 9.1. The Truth Table for an Exclusive Disjunction Between Two Propositions, A and B

A	B	A or else B, but not both
True	True	False
True	False	True
False	True	True
False	False	False

Table 9.2. The Fully Explicit Models and the Mental Models of the Possibilities Consistent With Assertions Based on the Principal Sentential Connectives

Sentences	Fully Explicit Models		Mental Models	
A and B:	A	B	A	B
Neither A nor B:	not-A	not-B	not-A	not-B
A or else B but not both:	A	not-B	A	B
	not-A	B		
A or B or both:	A	not-B	A	B
	not-A	B		
	A	B	A	B
If A, then B:	A	B	A	B
	not-A	B	...	
	not-A	not-B		
If, and only if A, then B:	A	B	A	B
	not-A	not-B	...	

Note. “Not” denotes the representation of negation, and “...” denotes an implicit model, that is, one with no explicit content.

(Premise 1)	(Premise 2)	
king	not-ace	not-king
not-king	ace	

Their conjunction depends on combining each model in one set with each model in the other set according to two rules:

- A contradiction between a pair of models yields the null model (akin to the empty set).
- Any other conjunction yields a model of each proposition in the two models.

The result of conjoining these two sets of models is therefore a single non-null model:

not-king ace

Since an inference is valid if its conclusion holds in all the models of the premises, the conclusion, *there is an ace*, follows validly from the premises. The two preceding rules can be used recursively to construct the models of compound premises containing multiple connectives.

Because infinitely many conclusions follow from any premises, computer programs for proving validity generally evaluate given conclusions. Human

reasoners, however, can draw conclusions for themselves. They normally abide by two constraints (Johnson-Laird & Byrne, 1991). First, they do not throw semantic information away by adding disjunctive alternatives to those to which the premises refer. For instance, given a single premise, *there is a king*, they never spontaneously conclude, *there is a king or there is an ace, or both*, even though the deduction is valid. Second, they draw novel conclusions that are parsimonious. For instance, they never draw a conclusion that merely forms a conjunction of all the premises, even though such a deduction is valid. Of course, human performance rapidly degrades with complex problems, but the goal of parsimony suggests that intelligent programs should draw conclusions that succinctly express all the information in the premises. The model theory yields an algorithm that draws such conclusions (Chapter 9, Johnson-Laird & Byrne, 1991).

Fully explicit models are simpler than truth tables, but they still place a heavy load on working memory. Mental models are still simpler, because they are limited by the third principle of the theory (Johnson-Laird, 2006):

- III. The principle of *truth*: Mental models represent only what is possible given the truth of assertions, not what is impossible, and each mental model represents

a clause in the premises only when the clause is true in the possibility that the model represents.

The simplest test of the principle is to ask naive individuals to list what is possible for a variety of assertions (Barrouillet & Lecas, 1999; Johnson-Laird & Savary, 1996). Given an exclusive disjunction, such as *There is a king in the hand or else there isn't an ace*, they list two possibilities corresponding to the mental models:

king	not-ace
------	---------

The first mental model does not represent that it is false that there isn't an ace in this possibility, that is, there *is* an ace; and the second mental model does not represent that it is false that there is a king in this possibility, that is, there *isn't* a king. Hence, people tend to neglect these cases. Readers might assume that the principle of truth is equivalent to the representation of the propositions *mentioned* in the premises. But this assumption yields the same models of clauses regardless of the connective relating them. A simple and approximate way to conceive the principle is therefore that it yields pared-down versions of fully explicit models, which in turn map into truth tables. As we will see, the principle of truth predicts a striking phenomenon in reasoning.

Individuals can make a mental footnote about what is false in a possibility, and these footnotes can be used to flesh out mental models into fully explicit models. But footnotes tend to be ephemeral. The most recent computer program implementing the model theory operates at two levels of expertise. At its more expert level, it uses *fully explicit* models, and so it makes no mistakes in reasoning. At its less expert level, it uses *mental* models without footnotes. Its representations for assertions based on the main sentential connectives are summarized in Table 9.2. The mental models of a conditional: *If there is a king, then there is an ace*, are as follows:

king	ace
...	

The ellipsis denotes an *implicit* model of the possibilities in which the if-clause of the conditional is false. In other words, there are alternatives to the possibility in which there is a king and an ace, but individuals tend not to think explicitly about what they are. If they retain the footnote about what is false, then they can flesh out the mental models into fully explicit models. The mental models of the biconditional: *If, and*

only if, there is a king, then there is an ace, as Table 9.2 shows, are identical to those for the conditional. What differs is that the footnote now conveys that *there is a king* and *there is an ace* are each false in the implicit model.

Inferences can be made with mental models using a procedure that builds a set of models for a premise and then updates them according to the other premises. From the premises:

There is a king or else there is an ace, but not both.
There is not a king.

the disjunction yields the mental models:

king	ace
------	-----

The categorical premise eliminates the first model, but it is compatible with the second model, yielding the valid conclusion: there is an ace. These procedures are summarized in Table 9.3.

The model theory was originally developed for so-called syllogisms, such as:

Some actuaries are businessmen.
All businessmen are conformists.
Therefore, some actuaries are conformists.

in which each assertion contains a quantifier, such as “all” or “some.” To account for the results, the theory was modified over the years (see, e.g., Johnson-Laird & Bara, 1984), but it became clear that in fact individuals develop different strategies for coping with syllogisms (Bucciarelli & Johnson-Laird, 1999). The model theory has also been extended to some sorts of inference from premises containing more than one quantifier (Johnson-Laird, Byrne, & Tabossi, 1989). Consider, for example, the following inference (from Cherubini & Johnson-Laird, 2004):

There are four persons: Ann, Bill, Cath, and Dave.
Everybody loves anyone who loves someone.
Ann loves Bill.
What follows?

Most people can envisage this model in which arrows denote the relation of *loving*:



Hence, they infer that everyone loves Ann. But if you ask them whether it follows that Cath loves Dave,

Table 9.3. The Procedures for Forming a Conjunction of a Pair of Models

1. For mental models only: The conjunction of a pair of models in which a proposition, B , in one model is not represented in the other model depends on the set of models of which this other model is a member. If B occurs in at least one of these models, then its absence in the current model is treated as negation, $\text{not-}B$. Otherwise, its absence is treated as its affirmation. The remaining mechanisms then apply.
2. For mental models only: The conjunction of an implicit model with a model representing propositions yields the null model by default (akin to the empty set), for example:

... and $B \wedge C$ yield nil.

But if none of the atomic propositions ($B \wedge C$) is represented in the set of models containing the implicit model, then the conjunction yields the model of the propositions, for example:

... and $B \wedge C$ yield $B \wedge C$.

Likewise:

... and ... yield ...

3. The conjunction of a pair of models containing respectively a proposition and its negation yield the null model, for example:

$A \wedge B$ and $\text{not-}A \wedge B$ yield nil

4. The conjunction of a null model with any model yields the null model, for example:

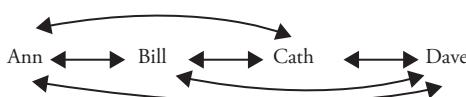
$A \wedge B$ and nil yield nil

5. The conjunction of a pair of fully explicit models free from contradiction update the second model with all the new propositions from the first model, for example:

$A \wedge B \wedge C$ and $A \wedge B$ yield $A \wedge B \wedge C$

Note. Each procedure is presented with an accompanying example. Only mental models can be implicit and therefore call for the first two procedures.

they tend to respond: “no.” They are mistaken, but the inference calls for using the quantified premise again. The result is this model (strictly speaking, all four persons love themselves, too):



It follows that Cath loves Dave, and people grasp its validity if it is demonstrated with such diagrams. No complete model theory exists for inferences based on quantifiers and connectives (cf. Bara, Bucciarelli, & Lombardo, 2001). But the main principles of the theory should apply: Mental models are iconic, they denote possibilities, and they represent only what is true.

Experimental Studies of Deductive Reasoning

Many experiments have corroborated the model theory, and the present section outlines the corroborations of five of its main predictions.

Prediction 1: The fewer the models needed for an inference, and the simpler they are, so the inference should take less time and be less prone to error. Fewer entities do improve inferences (e.g., Birney & Halford, 2002). Likewise, fewer models improve spatial and temporal reasoning (Byrne & Johnson-Laird, 1989; Carreiras & Santamaría, 1997; Schaeken, Johnson-Laird, & d'Ydewalle, 1996; Vandierendonck & De Vooght, 1997). Premises yielding one model of a spatial layout take less time to read than similar premises yielding multiple models, but the difference between two and three models is often so small that it is unlikely that reasoners construct all three models (Vandierendonck, De Vooght, Desimpelaere, & Dierckx, 1999). They may build a single model with one element represented as having two or more possible spatial locations.

Effects of number of models have been observed in comparing one sort of sentential connective with another and in examining batteries of such inferences

(Johnson-Laird & Byrne, 1991). As an illustration, consider a “double disjunction” (Bauer & Johnson-Laird, 1993):

Ann is in Alaska or else Beth is in Barbados, but not both.

Beth is in Barbados or else Cath is in Canada, but not both.

What follows?

Reasoners readily envisage the two possibilities compatible with the first premise, but it is harder to update them with those from the second premise. The solution yields the following two mental models, though real mental models represent spatial relations, not words or phrases:

Ann in Alaska

Cath in Canada

Beth in Barbados

These two models yield the conclusion: *Either Ann is in Alaska and Cath is in Canada, or else Beth is in Barbados*. An increase in number of models soon overloads working memory. The following inference defeats most people:

Ann is in Alaska or Beth is in Barbados, or both.

Beth is in Barbados or Cath is in Canada, or both.

What follows?

The premises yield five models, from which it follows: *Ann is in Alaska and Cath is in Canada, or Beth is in Barbados, or both*. When the order in which the premises are stated has the effect of reducing the number of models to be held in mind, reasoning improves (García-Madruga et al., 2001; Girotto, Mazzocco, & Tasso, 1997; Mackiewicz & Johnson-Laird, 2004).

Because one model is easier than many, an interaction occurs in *modal* reasoning, which concerns what is possible and what is necessary. It is easier to infer that a situation is possible (one model of the premises suffices as an example) than that it is *not* possible (all the models of the premises must be checked for a counterexample in which the situation occurs). In contrast, it is easier to infer that a situation is *not* necessary (one counterexample suffices) than that it is necessary (all the models of the premises must be checked as examples). The interaction occurs in both the accuracy and the speed of reasoning (Bell & Johnson-Laird, 1998).

Prediction 2: Reasoners should err as a result of overlooking models of premises. Given a double

disjunction (such as the ones mentioned earlier), the most frequent errors were conclusions consistent with just a single model of the premises (Bauer & Johnson-Laird, 1993). Likewise, many errors with syllogisms appear to arise because individuals consider only a single model of the premises (Bucciarelli & Johnson-Laird, 1999; Espino, Santamaría, & García-Madruga, 2000; Newstead & Griggs, 1999). Ormerod has proposed a “minimal completion” hypothesis according to which reasoners construct only the minimally necessary models (see Ormerod, Manktelow, & Jones, 1993; Richardson & Ormerod, 1997). Likewise, Sloutsky has postulated a process of “minimalization” in which reasoners tend to construct only single models for all connectives, thereby reducing them to conjunctions (Morris & Sloutsky, 2002; Sloutsky & Goldvarg, 1999). Certain assertions, however, do tend to elicit more than one model. As Byrne and her colleagues have shown (e.g., Byrne & McElaney, 2000; Byrne, 2002, 2005; Byrne & Tasso, 1999), counterfactual conditionals such as:

If the cable hadn't been faulty then the printer wouldn't have broken

tend to elicit models of both what is factually the case:

cable faulty printer broken

and what holds in a counterfactual possibility:

not cable faulty not printer broken

Prediction 3: Reasoners should be able to refute invalid inferences by envisaging counterexamples, that is, models of the premises that refute the putative conclusion. There is no guarantee that reasoners will find a counterexample, but, where they do succeed, they know that an inference is invalid (Barwise, 1993). The availability of a counterexample can suppress fallacious inferences from a conditional premise (Byrne, Espino, & Santamaría, 1999; Markovits, 1984; Vadeboncoeur & Markovits, 1999). Nevertheless, an alternative theory based on mental models has downplayed the role of counterexamples (Polk & Newell, 1995), and reasoners’ diagrams have sometimes failed to show their use (e.g., Newstead, Handley, & Buck, 1999). But when reasoners had to construct external models (Bucciarelli & Johnson-Laird, 1999), they used counterexamples (see also Neth & Johnson-Laird, 1999; Roberts, 2003).

There are two sorts of invalid conclusions. One sort is invalid because the conclusion is inconsistent with the premises, for example:

There is an apple or a banana on the table, or both.
There is a banana or else a cherry on the table, but not both.

Therefore, there isn't an apple on the table and there is a cherry on the table.

The premises have three fully explicit models:

apple	not-banana	cherry
not-apple	banana	not-cherry
apple	banana	not-cherry

The conclusion is inconsistent with the premises, because it conflicts with each of these models. Another sort of invalid conclusion, however, is consistent with the premises but does not follow from them. For instance, the conclusion *there is an apple on the table and there is a cherry*, holds in the first model, but it does not follow from the premises, because the other two models are counterexamples. Reasoners usually establish the invalidity of the first sort of conclusion by detecting its inconsistency with the premises, but they refute the second sort of conclusion with a counterexample (Johnson-Laird & Hasson, 2003). An experiment using functional magnetic resonance imaging examined reasoning based on numerical quantifiers, such as “at least five,” and mental arithmetic based on the same premises. Only a search for counterexamples in reasoning activated the right frontal pole, a region of the brain known to mediate inconsistencies (Kroger, Nystrom, Cohen, & Johnson-Laird, 2008).

Prediction 4: Reasoners should succumb to illusory inferences, which are compelling but invalid. They arise from the principle of truth, and its corollary that reasoners neglect what is false. Consider the problem:

Only one of the following assertions is true about a particular hand of cards:

There is a king in the hand or there is an ace, or both.

There is a queen in the hand or there is an ace, or both.

There is a jack in the hand or there is a ten, or both.

Is it possible that there is an ace in the hand?

Nearly everyone responds: “yes” (Goldvarg & Johnson-Laird, 2000). They grasp that the first assertion allows two possibilities in which an ace occurs. And so they infer that an ace is possible. In fact, it is impossible for an ace to be in the hand, because both of the first two assertions would then be true,

contrary to the rubric that only one of them is true. The inference is an illusion of possibility: Reasoners infer wrongly that a card is possible, because when they think about the truth of one assertion, they fail to think about the consequences of the falsity of other assertions. An analogous illusion of impossibility is created by replacing the two occurrences of *there is an ace* in the problem with *there is not an ace*. When the aforementioned premises were stated with the question:

Is it possible that there is a jack?

the participants nearly all responded, “yes.” They considered the third assertion, and its mental models showed that there could be a jack. But this time they were correct: The inference is valid. Hence, the focus on truth does not always lead to error, and experiments have compared illusions with similar control problems for which the neglect of falsity did not affect accuracy.

The computer program implementing the theory shows that illusory inferences should be sparse in the set of all possible inferences. But experiments have corroborated their occurrence in reasoning about possibilities in simple disjunctions (Khemlani & Johnson-Laird, 2009), probabilities, and causal and deontic relations (Buccicarelli & Johnson-Laird, 2005). Table 9.4 illustrates some typical illusions. Studies have used remedial procedures to reduce the illusions (e.g., Santamaría & Johnson-Laird, 2000). Yang taught participants to think explicitly about what is true and what is false. The difference between illusions and control problems vanished, but performance on the control problems fell from almost 100% correct to around 75% correct (Yang & Johnson-Laird, 2000). The principle of truth limits understanding, but it does so without participants realizing it. They were highly confident in their responses, usually no less so when they succumbed to an illusion than when they responded correctly to a control problem.

The rubric, “one of these assertions is true and one of them is false,” is equivalent to an exclusive disjunction between two assertions: *A or else B, but not both*. This usage leads to compelling illusions that seduce novices and experts alike, for example:

If there is a king then there is an ace or else if there isn't a king then there is an ace.

There is a king.

What follows?

Over 2,000 individuals have tackled this problem (see Johnson-Laird & Savary, 1999), and nearly

Table 9.4. Some Illusory Inferences in Abbreviated Form, With the Percentages of Illusory Responses

Premises Responses	Illusory Responses	Percentages of Illusory
1. If A then B or else B. A.	B.	100
2. Either A and B, or else C and D. A.	B.	87
3. If A then B or else if C then B. A and B.	Possibly both assertions are true.	98
4. A or else not both B and C. A and not B.	Possibly both assertions are true.	91
5. One true and one false: not-A or not-B, or neither. Not-C and not-B.	Possibly not-C and not-B.	85
6. Only one is true: At least some A are not B. No A are B.	Possibly No B are A.	95
7. If one is true so is the other: A or else not B. A.	A is more likely than B.	95
8. If one is true so is the other: A if and only if B. A.	A is equally likely as B.	90
9. Either A or else (B or else C).	A and B is not possible.	83
10. Either A, B, and C, or else C. A and not B.	The two assertions cannot both be true.	95

Notes. Each study examined other sorts of illusions and matched control problems. 1 is from Johnson-Laird & Savary (1999), 2 is from Walsh & Johnson-Laird (2004), 3 is from Johnson-Laird, Legrenzi, Girotto, & Legrenzi (2000), 4 is from Legrenzi, Girotto, & Johnson-Laird (2003), 5 is from Goldvarg & Johnson-Laird (2000), 6 is from Experiment 2, Yang & Johnson-Laird (2000), 7 and 8 are from Johnson-Laird & Savary (1996), 9 is from Khemlani & Johnson-Laird (2009), and 10 is from Byrne, Lotstein, & Johnson-Laird (in press).

everyone responded: *There is an ace*. The prediction of an illusion depends not on logic but on how other participants interpreted the relevant connectives in simple assertions. In particular, *or else* implies that one assertion that it connects can be true and the other assertion false; and if a conditional is false, then one possibility is that its if-clause is true and its then-clause is false. The illusion also occurs with the rubric *One of these assertions is true and one of them is false* applied to the two conditionals. The fact that the response is illusory is also borne out by the effect of various remedial procedures. Readers may suspect that the illusions arise from the artificiality of the problems, which never occur in real life, and which therefore confuse the participants. The problems may be artificial, though analogs do occur in real life (see Johnson-Laird & Savary, 1999), and artificiality fails to explain the correct responses to the controls or the high ratings of confidence in both illusory and control conclusions.

Prediction 5: Naive individuals should develop different reasoning strategies based on models. When they are tested in the laboratory, they start with only rough ideas of how to proceed. They can

reason, but not efficiently. With experience but no feedback about accuracy, they spontaneously develop various strategies (Schaeken, De Vooght, Vandierendonck, & d'Ydewalle, 1999). Deduction itself may be a master strategy (Evans, 2000), and people may resort to it more in Western cultures than in East Asian cultures (Peng & Nisbett, 1999). But deduction itself leads to different strategies (Van der Henst, Yang, & Johnson-Laird, 2002). Consider a problem in which each premise is *compound*, that is, contains a connective, and refers to different colored marbles that may be in a box:

There's a red if and only if there's a blue.
Either there's a blue or else there's a green, but not both.
There's a green if and only if there's a brown.
Does it follow that if there isn't red then there's a brown?

Some individuals develop a strategy based on suppositions. They say, for example:

Suppose there isn't a red. It follows from the first premise that isn't a blue. It follows from the second

premise that there's a green. The third premise then implies there's a brown. So, yes, the conclusion follows.

Some individuals construct a chain of conditionals leading from one clause in the conclusion to the other, for example: If there's a brown then there's a green, if there's a green then there isn't a blue, if there isn't a blue then there isn't a red. But most individuals develop a strategy in which they enumerate the different possibilities compatible with the premises. For example, they draw a horizontal line across the page and write down the two possibilities for the preceding premises:

red	blue
green	brown

When individuals are taught to use this strategy, as Victoria Bell showed in unpublished studies, their reasoning is both faster and more accurate. The nature of the premises and the conclusion can bias reasoners to adopt a predictable strategy, for example, conditional premises encourage the use of suppositions, whereas disjunctive premises encourage the enumeration of possibilities (Van der Henst et al., 2002).

Reasoners develop diverse strategies for relational reasoning (e.g., Goodwin & Johnson-Laird, 2005; Roberts, 2000), suppositional reasoning (e.g., Byrne & Handley, 1997), and reasoning with quantifiers (e.g., Buccirelli & Johnson-Laird, 1999). Granted the variety of strategies, there remains a robust effect: Inferences from one mental model are easier than those from multiple models (see also Espino, Santamaría, Meseguer, & Carreiras, 2000). Different strategies could reflect different mental representations (Stenning & Yule, 1997), but those so far discovered are all compatible with models. Individuals who have mastered logic could make a strategic use of formal rules. Given sufficient experience with a class of problems, individuals begin to notice some formal patterns.

Probabilistic Reasoning

Reasoning about probabilities is of two sorts. In *intensional* reasoning, individuals use heuristics to infer the probability of an event from some sort of evidence, such as the availability of information. In *extensional* reasoning, they infer the probability of an event from knowledge of the different ways in which it might occur. This distinction is due to Kahneman and Tversky, who together pioneered the investigation of heuristics in intensional reasoning (Kahneman, Slovic, & Tversky, 1982). Studies of

extensional reasoning focused at first on "Bayesian" reasoning in which participants try to infer a conditional probability of an event from various other probabilities. These studies offered no account of the mental processes of extensional reasoning, but the model theory filled the gap (Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999).

Mental models represent the extensions of assertions, that is, the possibilities to which they refer. The theory postulates the following:

The principle of *equiprobability*: Each mental model represents an equiprobable possibility unless there are reasons to the contrary.

The probability of an event accordingly depends on the proportion of models in which it occurs. The theory also allows that models can be tagged with numerals denoting probabilities or frequencies of occurrence, and that individuals can carry out primitive arithmetical operations on these numerals. Shimojo and Ichikawa (1989) and Falk (1992) proposed similar principles for Bayesian reasoning. The present account differs from theirs in that it assigns equiprobability, not to actual events but to mental models. And equiprobability applies only by default. Classical probability theory used an analogous principle of "indifference" (Hacking, 1975), but it is problematic because it applies to events per se rather than their mental representations.

Consider a simple problem from Johnson-Laird et al. (1999):

In the box, there is a green ball or a blue ball or both. What is the probability that both the green and the blue ball are there?

The premise elicits the mental models:

green	blue
green	blue

Naive reasoners follow the equiprobability principle, and infer the answer: 1/3. An experiment corroborated this and other predictions for the connectives in Table 9.2.

Conditional probabilities are on the borderline of naive competence. They are difficult because individuals need to consider several models. Here is a typical Bayesian problem:

The patient's PSA score is high. If he doesn't have prostate cancer, the chances of such a value is 1 in 1,000. Is he likely to have prostate cancer?

Many people respond: yes. But they are wrong. The model theory predicts the error: Individuals represent the conditional probability in the problem as one explicit model and one implicit model tagged with their respective chances:

not prostate cancer	high PSA	1
	...	999

The converse conditional probability has the same mental models, and so people are likely to assume that if the patient has a high PSA the chances are only 1 in 1,000 that he does not have prostate cancer, that is, given that the patient has a high PSA, the probability that he has prostate cancer is 999/1,000. To reason correctly, however, individuals must envisage the complete partition of possibilities and their chances. But the problem fails to provide enough information to do so. It yields only:

not prostate cancer	high PSA	1
not prostate cancer	not high PSA	999
prostate cancer	high PSA	?
prostate cancer	not high PSA	?

There are various ways to provide the missing information. One way is to give the base rate of prostate cancer, which can be used with Bayes' theorem in the probability calculus to infer the answer. But the theorem and its computations are beyond naive individuals (Kahneman & Tversky, 1973). The model theory postulates an alternative method:

The *subset principle*: Given a complete partition, individuals infer the conditional probability of *B* given *A* by computing the proportion corresponding to the subset of *B* that is *A*.
(Johnson-Laird et al., 1999)

If models are tagged with their absolute frequencies or chances, then this conditional probability is the value for *A and B* divided by the sum for all the models containing *B*. Suppose that the complete partition for the problem is as follows:

not prostate cancer	high PSA	1
not prostate cancer	not high PSA	999
prostate cancer	high PSA	2
prostate cancer	not high PSA	0

The probability of prostate cancer given a high PSA is the subset in which both occur (row 3) as a proportion of the cases of a high PSA (the sum of

rows 1 and 3), that is, 2/3. This probability is high but a long way from 999/1,000.

Evolutionary psychologists postulate that natural selection led to an innate “module” in the mind that makes Bayesian inferences from naturally occurring frequencies. It follows that naive reasoners should fail the patient problem because it is about a unique event (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995). In contrast, as the model theory predicts, individuals cope with problems about unique (or repeated) events provided they can use the subset principle and the arithmetic is easy (Girotto & Gonzalez, 2001). Even infants are able to grasp the basics of probabilistic reasoning (Téglás, Girotto, Gonzalez, & Bonatti, 2007).

The model theory dispels some common misconceptions about probabilistic reasoning. It is *not* always inductive. Extensional reasoning can be deductively valid, and it need not depend on a tacit knowledge of the probability calculus. It is not always correct, because it can yield illusions (see items 7 and 8 in Table 9.4).

Induction and Models

Induction, Modulation, and Meaning

Induction is part of everyday thinking, but Popper (1972) argued that it is not part of scientific thinking. He claimed that science is based on explanatory conjectures, which observations serve potentially only to falsify. Some scientists agree (e.g., Deutsch, 1997, p. 159). But many astronomical, meteorological, and medical observations are not tests of hypotheses (see the chapters in Part VII). And everyone makes inductions in daily life. For instance, when the starter will not turn over the engine, your immediate thought is that the battery is dead. You are likely to be right, but there is no guarantee. Likewise, when the car ferry, *Herald of Free Enterprise*, sailed from Zeebrugge on March 6, 1987, its master made the plausible induction that the bow doors had been closed. They had always been closed in the past, and he had no evidence to the contrary. But they had not been closed, the vessel capsized and sank, and many people drowned. Induction is a common but risky business.

The textbook definition of *induction*—alas, all too common—is that it is an inference from the particular to the general. Such arguments are indeed inductions, but many other inductions such as the two earlier examples are inferences from the particular to the particular. That is why the Introduction to this chapter offered a more comprehensive definition: Induction

is a process that increases semantic information. As an example, consider again the inference:

The starter won't turn.
Therefore, the battery is dead.

Like all inductions, it depends on knowledge and in particular on the true conditional:

If the battery is dead, then the starter won't turn.

The conditional refers to three possibilities:

battery dead	not starter turn
not battery dead	not starter turn
not battery dead	starter turn

The premise of the induction eliminates the third possibility, but the conclusion goes beyond the given information because it eliminates the second possibility, too. The availability of the first model yields an intensional inference of a high probability, but the conclusion may be false. Inductions are risky because they increase semantic information.

Inductions depend on knowledge. As Kahneman and Tversky have shown, various heuristics constrain the use of knowledge in inductions. The present account presupposes these heuristics, but it examines the role of models in induction. The model theory gives a “dual-process” account of inferences: *Intuitive* reasoning has no access to working memory, and so it is incapable of recursion—loops of operations such as those in the earlier inference about Cath loving Dave—and incapable of representing alternative possibilities (Johnson-Laird, 1983, Chapter 9, 2006). Intuitive inductions are accordingly dependent on what is known as “System 1,” which is nonrecursive and therefore rapid. It is unconscious and therefore involuntary. And it yields only a single mental model of the situation. Other inductions are *explicit*: They are slow, voluntary, and conscious. Dual-process theories are long-standing (e.g., Evans & Over, 1996; Johnson-Laird & Wason, 1977, p. 341; Sloman, 1996; Stanovich, 1999; see Evans, Chapter 8; Stanovich, Chapter 22). What the model theory provides is a distinction in computational power: System 1 is computationally weak because it makes no use of working memory, whereas System 2 uses working memory and is therefore capable of recursion and the representation of alternative possibilities.

One sort of intuitive induction affects the interpretation of sentences and has subsequent effects on

reasoning. It is known as modulation (Johnson-Laird & Byrne, 2002):

The principle of *modulation*: the meanings of clauses, coreferential links between them, general knowledge, and knowledge of context, can modulate the core meanings of sentential connectives (which are shown in Table 9.2).

One effect is to block the construction of models of possibilities. Consider the conditional:

If it's a game, then it's not soccer.

It is compatible with only two possibilities:

game	not-soccer
not-game	not-soccer

The knowledge that soccer is a game blocks the construction of the possibility in which “it” refers to something that is not a game, but that is soccer. Hence, the conditional alone yields the valid inference: *It's not soccer*. Depending on the contents and the grammatical form of a conditional, modulation yields 10 distinct interpretations of conditionals depending solely on which possibilities it blocks (Johnson-Laird & Byrne, 2002). But, as we will see, still other interpretations of conditionals can occur. A corollary, however, is that no form of inference, such as: *if A then C; not C; therefore, not-A*, is guaranteed to yield a valid deduction regardless of its contents. The validity of inferences can be decided only on a case-by-case basis: Those inferences in which the conclusion must be true if the premises are true are valid. They are valid, not in virtue of their “logical form,” but in virtue of content. Hence, the model theory makes no use of logical form (Byrne & Johnson-Laird, 2009).

Experiments have shown that modulation occurs and affects the different sets of possibilities that participants list for assertions, which in turn predict the conclusions that participants in other experiments tend to draw (Quelhas, Johnson-Laird, & Juhos, 2010). Consider these two conditionals (translated from the Portuguese):

If the dish is kidney beans, then its basis is beans.
If the dish is made of meat, then it can be
Portuguese stew.

A subtle difference exists between their interpretations. When individuals evaluate possibilities in relation to the first sort of conditional (*If A then B*), they judge that it is possible that the dish isn't kidney

beans but its basis is beans (*not-A and B*), but that it is not possible that the dish is kidney beans and its basis isn't beans (*A and not-B*). But they make the opposite evaluations for the second conditional: It is impossible that the dish isn't made of meat and is Portuguese stew (*not-A and B*), but it is possible that that dish is made of meat and is not Portuguese stew (*A and not-B*). These interpretations, and those for other sorts of conditional, lead to quite different but predictable patterns of inference. For example, given an assertion of the if-clause in the conditional about the beans, *A*, reasoners infer that *B* follows; but given a denial of the if-clause, *not-A*, they refrain from inferring that *not-C* follows. They make the opposite pattern of inferences from the conditional about the stew.

Models can represent spatial and temporal information (Byrne & Johnson-Laird, 1989; Schaeken et al., 1996). And another effect of modulation is to add such information to models. For instance, the assertion, *If she put a book on the shelf, then it fell off*, elicits the temporal relation that the book was put on the shelf before it fell, and the spatial relation that the book ended up below the shelf (Johnson-Laird & Byrne, 2002). The effects of temporal modulation were borne out in a study in which individuals drew conclusions corroborating the predicted temporal relations, which in some cases were that the event in the if-clause occurred first, and in other cases were that the event in the then-clause occurred first (Quelhas et al., 2010). Modulation is rapid and automatic, and it affects comprehension and reasoning (Johnson-Laird & Byrne, 2002; Newstead, Ellis, Evans, & Dennis, 1997; Quelhas et al., 2010). In logic, connectives such as conditionals and disjunctions are *truth functional*, and so the truth-value of a sentence in which they occur can be determined solely from a knowledge of the truth-values of the clauses they interconnect (see Table 9.1). But in natural language connectives are not truth functional: The process of their interpretation has to check whether content and context modulate their meaning.

The Detection of Inconsistencies

When the premises and conclusion of an inference refer to disjoint possibilities, they are inconsistent with one another. But inconsistency is a more general phenomenon: A set of assertions can be inconsistent, even though any proper subset of them is consistent. For example, these three

assertions are inconsistent, that is, they cannot all be true at the same time:

If there is a square on the board, then there is a triangle.

If there is a triangle, then there is a circle.

There is a square on the board and there isn't a circle.

Yet any pair of them is consistent. This principle applies in general: A set of n propositions can be inconsistent even though any $n-1$ of them is consistent. In principle, 100 propositions, which can each be true or false, yield 2^{100} cases. To determine whether the propositions are consistent may call for an examination of all these cases to see whether there is one in which all the propositions are true. Even if you could examine each possibility in a millionth of a second, it would still take you much longer than the universe has existed to examine them all. In other words, the evaluation of a set of assertions as consistent or inconsistent is computationally intractable (Cook, 1971). As the number of possibilities increases, it soon ceases to be feasible to make an exhaustive check.

Some psychologists argue that logically untrained individuals are incapable of deductive reasoning, or reluctant to use it, and instead use various heuristics (e.g., Oaksford & Chater, 2007; Wetherick & Gilhooly, 1995). But if untrained individuals are able to detect inconsistencies, they are in effect making deductions. The two tasks seem different, but they are merely opposite sides of one another. This relation is exploited in one method of logic: to prove that a conclusion follows validly, you add its negation to the premises and show that the resulting set of propositions is inconsistent (see, e.g., Jeffrey, 1981). In fact, untrained individuals are deductively competent because they can detect inconsistencies. The evidence implies that they do so by searching for a model in which all the propositions hold. If they find such a model, they evaluate the propositions as consistent—they could all be true at the same time; otherwise, they evaluate the propositions as inconsistent.

The theory predicts that when the first model that individuals are likely to envisage satisfies a set of assertions, then the task should be easier than when it does not, and they have to search for an alternative model. Consider, for instance, these assertions about what is on top of a table:

If there isn't an apple, then there is a banana.

If there is a banana, then there is a cherry.

There isn't an apple and there is a cherry.

The explicit mental model of the first assertion (see Table 9.2) is one that satisfies its two clauses:

not-apple banana

The second assertion holds in this model and updates it:

not-apple banana cherry

The third assertion holds in this model, and so reasoners should respond that the assertions are consistent. A contrasting set of assertions is as follows:

There is an apple or there is a banana.

There isn't a banana or there is a cherry.

There isn't an apple and there is a cherry.

Individuals are likely to start with a model of the first clause of the first assertion:

apple

They may continue to update it only to discover that the third assertion is not consistent with it. They now have to start over with an alternative model of the first assertion:

banana

This model refutes the first clause of the second assertion, and so its second clause holds:

banana cherry

The third assertion holds in this possibility, and so individuals can now respond that the three assertions can all be true at the same time. This predicted difference in performance occurred in an experiment, which counterbalanced the two conditions over conditionals and disjunctions (Johnson-Laird, Legrenzi, Girotto, & Legrenzi, 2000).

Another line of evidence corroborating the use of models showed that individuals succumb to illusions of consistency and of inconsistency. Consider the exclusive disjunction:

If there's an apple then there's a banana, or else if there's a cherry then there's a banana.

Its mental models are as follows:

apple banana
 banana cherry

They imply that a second assertion:

There is an apple and a banana

is consistent with the first one. In contrast, the fully explicit models of the disjunction take into account that when one conditional is true, the other conditional is false. And individuals take the falsity of a conditional, such as *if there's an apple, then there's a banana* to mean that there can be an apple without a banana (e.g., Barres & Johnson-Laird, 2003). When the first conditional is true, the second conditional is false and so there is a cherry and not a banana, and this case can occur when there isn't an apple and isn't a banana according to the truth of the first conditional. Conversely, when the second conditional is true, the first conditional is false, and so the fully explicit models of the disjunction are as follows:

apple	not-banana	not-cherry
not-apple	not-banana	cherry

These models show that the correct response is that the previous two assertions cannot both be true: They are inconsistent with one another. Similar discrepancies yield illusions of inconsistency: Individuals infer that assertions are inconsistent when in fact they are consistent. Nevertheless, they respond correctly to control problems in which the neglect of falsity has no effect on accuracy. For example, they correctly evaluate these assertions as consistent:

If there's an apple then there's a banana, or else if there's a cherry then there's a banana.

There is an apple and not a cherry.

The second assertion holds in the mental models and the fully explicit models of the first assertion. Experiments have shown that individuals tend to err with illusions but to perform correctly with control problems (Johnson-Laird, et al., 2000; Legrenzi, Girotto, & Johnson-Laird, 2003).

Abduction, Causation, and the Creation of Explanations

Induction is the use of knowledge to increase semantic information: Possibilities are eliminated either by adding elements to a mental model or by eliminating a mental model altogether. Induction can yield general descriptions and scientific laws. For example, Kepler's first law describes the orbits of the planets as ellipses with the sun at one of the two foci, and it was an induction from Tycho Brahe's astronomical observations. Scientific theories explain such laws in terms of more fundamental considerations, for example, the general theory of

relativity explains planetary orbits as the result of the sun's mass curving space-time. Peirce (1955) called reasoning that leads to such explanations *abduction*. In terms of the five categories of the Introduction, abduction is creative when it overrides data or beliefs.

Consider this problem (from Lee & Johnson-Laird, 2006):

If a pilot falls from a plane without a parachute, the pilot dies. This pilot did not die, however. Why not?

Many people respond with putative explanations, such as: *The pilot fell into a deep snowdrift*. Others instead draw a logically valid conclusion: *The pilot did not fall from the plane without a parachute*. The human propensity to explain is extraordinary, and it goes far beyond what can currently be simulated in any computer program. Tony Anderson and the present author illustrated this point when they asked participants, in effect, to explain the inexplicable. The participants received pairs of sentences selected at random from separate stories, also selected at random, for example:

John made his way to a shop which sold TV sets.
Celia had recently had her ears pierced.

In another condition of the experiment, the sentences were edited to make them coreferential:

Celia made her way to a shop which sold TV sets.
She had recently had her ears pierced.

The participants' task was to explain what was going on. They readily went beyond the given information to account for what was happening. They proposed, for example, that Celia was getting reception in her earrings and wanted the TV shop to investigate, that she wanted to see some new earrings on closed circuit TV, that she had won a bet by having her ears pierced and was spending the money on a TV set. Only rarely were the participants stumped for an explanation, and they were almost as ingenious with the sentences that were not coreferential.

Abduction depends on knowledge, especially of causal relations, which according to the model theory refer to temporally ordered sets of possibilities (Goldvarg & Johnson-Laird, 2001; cf. Buehner & Cheng, Chapter 12). An assertion of the form *C causes E* refers to three possibilities:

C	E
not-C	E
not-C	not-E

with the temporal constraint that *E* cannot precede *C*. An "enabling" assertion of the form *C allows E* refers to the three possibilities with the same temporal constraint:

C	E
C	not-E
not-C	not-E

This account accordingly distinguishes between the meaning of causes and of enabling conditions (pace, e.g., Einhorn & Hogarth, 1986; Hart & Honoré, 1985; Mill, 1874). It also treats the meaning of causal assertions as deterministic rather than probabilistic (cf. Cheng, 1997; Suppes, 1970). Experiments support both these claims: Participants listed the possibilities mentioned earlier, and they rejected other cases as impossible, contrary to probabilistic accounts (Goldvarg & Johnson-Laird, 2001). However, when individuals *induce* a causal relation from a series of observations, they may be influenced by the relative frequencies of various sorts of observation (Perales & Shanks, 2008; but cf. Lagnado, Waldmann, Hagmayer, & Sloman, 2007). Yet the meaning of a causal relation seems to be deterministic, because individuals claim that a single counterexample suffices to refute a general claim, such as: *contact between these two sorts of substance causes an explosion to occur* (Frosch & Johnson-Laird, 2009).

Given the cause from a causal relation, there is only one possible effect, as the models show; but given the effect, there is more than one possible cause. Exceptions do occur (Cummins, Lubart, Alksnis, & Rist, 1991; Markovits, 1984), but the principle holds in general. It explains why inferences from causes to effects are more plausible than inferences from effects to causes. As Tversky and Kahneman (1982) showed, conditionals in which the if-clause is a cause, such as:

A girl has blue eyes if her mother has blue eyes,

are judged as more probable than conditionals in which the if-clause is an effect:

The mother has blue eyes if her daughter has blue eyes.

According to the model theory, when individuals discover inconsistencies, their immediate task is, not to revise their beliefs, but rather to construct a causal

model that resolves the inconsistency. Consider, for example, the scenario:

If the trigger is pulled, then the pistol will fire. The trigger is pulled, but the pistol does not fire. Why not?

Given 20 different scenarios of this form, most explanations were causal claims that repudiated the conditional (Johnson-Laird, Girotto, & Legrenzi, 2004). In a further experiment with the scenarios, the participants rated the statements of a cause and its effect as the most probable explanations, for example:

A prudent person had unloaded the pistol and there were no bullets in the chamber.

The cause alone was rated as less probable, but as more probable than the effect alone, which in turn was rated as more probable than an explanation that repudiated the categorical premise, for example:

The trigger wasn't really pulled.

The greater probability assigned to the conjunction of the cause and effect than to either of its clauses is an instance of the “conjunction” fallacy in which a conjunction is in error judged to be more probable than its constituents (Tversky & Kahneman, 1983), and it suggests that models of causes and their effects are a staple in long-term memory.

Abductions that resolve inconsistencies have been implemented in a computer program that uses a knowledge base to create causal explanations (Johnson-Laird et al., 2004). Given the preceding example, the program constructs the mental models of the conditional:

The conjunction of the categorical assertion yields:

The fact that the pistol did not fire is inconsistent with this model. The theory predicts that individuals should tend to abandon their belief in the conditional premise, because its one explicit mental model conflicts with the fact that the pistol did not fire (see Girotto, Johnson-Laird, Legrenzi, & Sonino, 2000, for corroborating evidence). Nevertheless, the conditional is a useful idealization, and so the

program uses it to construct a counterfactual set of possibilities:

trigger pulled	not pistol fires	[the model of the facts]
trigger pulled	pistol fires	[the models of counterfactual possibilities]

People know that a pistol without bullets does not fire, and so the program has in its knowledge base the fully explicit models:

not bullets in pistol	not pistol fires
bullets in pistol	not pistol fires
bullets in pistol	pistol fires

The model of the facts triggers the first possibility in this set, which modulates the model of the facts to create a possibility:

not bullets in trigger not pistol fires
pistol pulled [possibility]

The new proposition in this model triggers a causal antecedent from another set of models in the knowledge base, which completes the process:

person	not bullets	trigger	not
empties	in pistol	pulled	pistol fires
pistol			
not person	bullets in	trigger	pistol
empties	pistol	pulled	fires
pistol			

The first of the three models above denotes a possibility that explains the inconsistency: A person emptied the pistol and so it had no bullets. The remaining models denote counterfactual possibilities, which yield the counterfactual claim: If the person hadn't emptied the pistol, then it would have had bullets, and...it would have fired. The fact that the pistol did not fire has been used to reject the conditional premise, and available knowledge has been used to create an explanation and to modulate the conditional premise into a counterfactual. There are, of course, other possible explanations for the failure of the pistol to fire. And so the program has a set of alternatives from which it makes an arbitrary choice.

In sum, reasoners can resolve inconsistencies between incontrovertible evidence and the consequences of their beliefs. They use their available knowledge—in the form of explicit models—to try to create a causal scenario that makes sense of the facts. They may resolve the inconsistency, create an erroneous account, or fail to construct any explanation whatsoever.

Conclusions and Future Directions

The concept of a mental model has its antecedents in the 19th century. The present theory was developed in the 20th century. In its application to deduction, as Peirce anticipated, if a conclusion holds in all the models of the premises, it is necessary given the premises. If it holds in a proportion of the models, then, granted that they are equiprobable, its probability is equal to that proportion. If it holds in at least one model, then it is possible. The theory also applies to inductive reasoning—both the rapid implicit inferences of System 1 that underlie comprehension and the deliberate inferences of System 2 that yield generalizations. And it offers an account of the creation of explanations that resolve inconsistencies. But if mental models underlie all thinking with a propositional content, then the theory remains incomplete. It is currently under intensive development and intensive scrutiny. Indeed, there are distinguishable variants of the theory itself (see, e.g., Evans, 1993; Ormerod et al., 1993; Polk & Newell, 1995; Schroyens & Schaeken, 2003; Schroyens, Schaeken, & Dieussaert, 2008). The three most urgent questions to be answered in the 21st century are as follows:

1. What role do mental models play in problem solving?
2. In what ways, if any, do *intensional* estimates of probabilities depend on mental models?
3. Current programs simulate the operations of System 2 in deliberate reasoning, but what is the nature of the algorithm underlying the use of models in the intuitive thinking of System 1?

Acknowledgments

This chapter was made possible by a grant from the National Science Foundation (grant SES 0844851 to study deductive and probabilistic reasoning). The author is grateful to the editors, to the community of reasoning researchers, and to his colleagues, collaborators, and students—many of whose names are to be found in the References.

References

- Bara, B. G., Buccarelli, M., & Lombardo, V. (2001). Model theory of deduction: A unified computational approach. *Cognitive Science*, 25, 839–901.
- Bar-Hillel, Y., & Carnap, R. (1964). An outline of a theory of semantic information. In Y. Bar-Hillel (Ed.), *Language and information processing* (pp. 221–274). Reading, MA: Addison-Wesley.
- Barres, P., & Johnson-Laird, P. N. (2003). On imagining what is true (and what is false). *Thinking & Reasoning*, 9, 1–42.
- Barrouillet, P., & Lecas, J-F. (1999). Mental models in conditional reasoning and working memory. *Thinking & Reasoning*, 5, 289–302.
- Barwise, J. (1993). Everyday reasoning and logical inference. *Behavioral and Brain Sciences*, 16, 337–338.
- Bauer, M. I., & Johnson-Laird, P. N. (1993). How diagrams can improve reasoning. *Psychological Science*, 4, 372–378.
- Bell, V., & Johnson-Laird, P. N. (1998). A model theory of modal reasoning. *Cognitive Science*, 22, 25–51.
- Birney, D., & Halford, G. S. (2002). Cognitive complexity of suppositional reasoning: An application of relational complexity to the knight-knave task. *Thinking & Reasoning*, 8, 109–134.
- Braine, M. D. S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review*, 85, 1–21.
- Bransford, J. D., Barclay, J. R., & Franks, J. J. (1972). Sentence memory: A constructive versus an interpretive approach. *Cognitive Psychology*, 3, 193–209.
- Buccarelli, M., & Johnson-Laird, P. N. (1999). Strategies in syllogistic reasoning. *Cognitive Science*, 23, 247–303.
- Buccarelli, M., & Johnson-Laird, P. N. (2005). Naïve deontics: A theory of meaning, representation, and reasoning. *Cognitive Psychology*, 50, 159–193.
- Byrne, R. M. J. (2002). Mental models and counterfactual thoughts about what might have been. *Trends in Cognitive Sciences*, 6, 426–431.
- Byrne, R. M. J. (2005). *The rational imagination: How people create alternatives to reality*. Cambridge, MA: MIT.
- Byrne, R. M. J., Espino, O., & Santamaría, C. (1999). Counterexamples and the suppression of inferences. *Journal of Memory and Language*, 40, 347–373.
- Byrne, R. M. J., & Handley, S. J. (1997). Reasoning strategies for suppositional deductions. *Cognition*, 62, 1–49.
- Byrne, R. M. J., & Johnson-Laird, P. N. (1989). Spatial reasoning. *Journal of Memory and Language*, 28, 564–575.
- Byrne, R. M. J., & Johnson-Laird, P. N. (2009). ‘If’ and the problems of conditional reasoning. *Trends in Cognitive Sciences*, 13, 282–286.
- Byrne, R. M. J., Lotstein, M., & Johnson-Laird, P. N. (in press). Disjunctions: A theory of meaning, pragmatics, and inference.
- Byrne, R. M. J., & McEleney, A. (2000). Counterfactual thinking about actions and failures to act. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 1318–1331.
- Byrne, R. M. J., & Tasso, A. (1999). Deductive reasoning with factual, possible, and counterfactual conditionals. *Memory & Cognition*, 27, 726–740.
- Carreiras, M., & Santamaría, C. (1997). Reasoning about relations: Spatial and nonspatial problems. *Thinking & Reasoning*, 3, 191–208.

- Cheng, P.W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, 104, 367–405.
- Cherubini, P., & Johnson-Laird, P.N. (2004). Does everyone love everyone? The psychology of iterative reasoning. *Thinking & Reasoning*, 10, 31–53.
- Cook, S. A. (1971). The complexity of theorem proving procedures. *Proceedings of the Third Annual Association of Computing Machinery Symposium on the Theory of Computing*, 151–158.
- Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, 58, 1–73.
- Craik, K. (1943). *The nature of explanation*. Cambridge, England: Cambridge University Press.
- Cummins, D. D., Lubart, T., Alksnis, O., & Rist, R. (1991). Conditional reasoning and causation. *Memory & Cognition*, 19, 274–282.
- de Kleer, J. (1977). Multiple representations of knowledge in a mechanics problem-solver. *International Joint Conference on Artificial Intelligence*, 299–304.
- Deutsch, D. (1997). *The fabric of reality: The science of parallel universes—and its implications*. New York: Penguin Books.
- Ehrlich, K. (1996). Applied mental models in human-computer interaction. In J. Oakhill & A. Garnham (Eds.), *Mental models in cognitive science* (pp. 223–245). Mahwah, NJ: Erlbaum.
- Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, 99, 3–19.
- Espino, O., Santamaría, C., Meseguer, E., & Carreiras, M. (2000). Eye movements during syllogistic reasoning. In J. A. García-Madruga, N. Carriero, & M. J. González-Labra (Eds.), *Mental models in reasoning* (pp. 179–188). Madrid, Spain: Universidad Nacional de Educación a Distancia.
- Espino, O., Santamaría, C., & García-Madruga, J. A. (2000). Activation of end terms in syllogistic reasoning. *Thinking & Reasoning*, 6, 67–89.
- Evans, J. St. B. T. (1993). The mental model theory of conditional reasoning: Critical appraisal and revision. *Cognition*, 48, 1–20.
- Evans, J. St. B. T. (2000). What could and could not be a strategy in reasoning. In W. S. Schaeken, G. De Vooght, A. Vandierendonck, & G. d'Ydewalle (Eds.), *Deductive reasoning and strategies* (pp. 1–22). Mahwah, NJ: Erlbaum.
- Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, England: Psychology Press.
- Falk, R. (1992). A closer look at the probabilities of the notorious three prisoners. *Cognition*, 43, 197–223.
- Frosch, C. A., & Johnson-Laird, P. N. (2009). Is causation probabilistic? *Proceedings of 31st Annual Conference of Cognitive Science Society*, 195–200.
- García-Madruga, J. A., Moreno, S., Carriero, N., Gutiérrez, F., & Johnson-Laird, P. N. (2001). Are conjunctive inferences easier than disjunctive inferences? A comparison of rules and models. *Quarterly Journal of Experimental Psychology*, 54A, 613–632.
- Garnham, A. (1987). *Mental models as representations of discourse and text*. Chichester, England: Ellis Horwood.
- Garnham, A. (2001). *Mental models and the interpretation of anaphora*. Hove, England: Psychology Press.
- Garnham, A., & Oakhill, J. V. (1996). The mental models theory of language comprehension. In B. K. Britton & A. C. Graesser (Eds.), *Models of understanding text* (pp. 313–339). Hillsdale, NJ: Erlbaum.
- Gentner, D., & Gentner, D. R. (1983). Flowing waters or teeming crowds: Mental models of electricity. In D. Gentner & A. L. Stevens (Eds.), *Mental models*. Hillsdale, NJ: Erlbaum.
- Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Erlbaum.
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency format. *Psychological Review*, 102, 684–704.
- Girotto, V., & Gonzalez, M. (2001). Solving probabilistic and statistical problems: A matter of question form and information structure. *Cognition*, 78, 247–276.
- Girotto, V., Johnson-Laird, P. N., Legrenzi, P., & Sonino, M. (2000). Reasoning to consistency: How people resolve logical inconsistencies. In J. A. García-Madruga, N. Carriero, & M. González-Labra (Eds.), *Mental models in reasoning* (pp. 83–97). Madrid, Spain: Universidad Nacional de Educación a Distancia.
- Girotto, V., Mazzocco, A., & Tasso, A. (1997). The effect of premise order in conditional reasoning: A test of the mental model theory. *Cognition*, 63, 1–28.
- Glenberg, A. M., Meyer, M., & Lindem, K. (1987). Mental models contribute to foregrounding during text comprehension. *Journal of Memory and Language*, 26, 69–83.
- Goldvarg, Y., & Johnson-Laird, P. N. (2000). Illusions in modal reasoning. *Memory & Cognition*, 28, 282–294.
- Goldvarg, Y., & Johnson-Laird, P. N. (2001). Naïve causality: A mental model theory of causal meaning and reasoning. *Cognitive Science*, 25, 565–610.
- Goodwin, G., & Johnson-Laird, P. N. (2005). Reasoning about relations. *Psychological Review*, 112, 468–493.
- Hacking, I. (1975). *The emergence of probability*. Cambridge, England: Cambridge University Press.
- Halford, G. S. (1993). *Children's understanding: The development of mental models*. Hillsdale, NJ: Erlbaum.
- Hart, H. L. A., & Honoré, A. M. (1985). *Causation in the law*. (2nd ed.). Oxford, England: Clarendon Press.
- Hayes, P. J. (1979). Naïve physics I – ontology for liquids. Mimeo, Centre pour les études Semantiques et Cognitives, Geneva, 1979. In J. Hobbs & R. Moore (Eds.), *Formal theories of the commonsense world*. Hillsdale, NJ: Erlbaum.
- Hegarty, M. (1992). Mental animation: Inferring motion from static diagrams of mechanical systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 1084–1102.
- Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press.
- Jeffrey, R. (1981). *Formal logic: Its scope and limits*. (2nd ed.). New York: McGraw-Hill.
- Johnson-Laird, P. N. (1970). The perception and memory of sentences. In J. Lyons (Ed.), *New horizons in linguistics* (pp. 261–270). Harmondsworth, England: Penguin Books.
- Johnson-Laird, P. N. (1975). Models of deduction. In R. Falmagne (Ed.), *Reasoning: Representation and process* (pp. 7–54). Springdale, NJ: Erlbaum.
- Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference and consciousness*. Cambridge, England: Cambridge University Press; Cambridge, MA: Harvard University Press.
- Johnson-Laird, P. N. (2002). Peirce, logic diagrams, and the elementary operations of reasoning. *Thinking & Reasoning*, 8, 69–95.

- Johnson-Laird, P. N. (2004). The history of mental models. In K. Manktelow (Ed.), *Psychology of reasoning: Theoretical and historical perspectives* (pp. 179–212). London: Psychology Press.
- Johnson-Laird, P. N. (2006). *How we reason*. New York: Oxford University Press.
- Johnson-Laird, P. N., & Bara, B. G. (1984). Syllogistic inference. *Cognition*, 16, 1–61.
- Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Erlbaum.
- Johnson-Laird, P. N., & Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review*, 109, 646–678.
- Johnson-Laird, P. N., Byrne, R. M. J., & Tabossi, P. (1989). Reasoning by model: The case of multiple quantification. *Psychological Review*, 96, 658–673.
- Johnson-Laird, P. N., & Hasson, U. (2003). Counterexamples in sentential reasoning. *Memory & Cognition*, 31, 1105–1113.
- Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review*, 111, 640–661.
- Johnson-Laird, P. N., Legrenzi, P., Girotto, P., & Legrenzi, M.S. (2000). Illusions in reasoning about consistency. *Science*, 288, 531–532.
- Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M., & Caverni, J.-P. (1999). Naïve probability: A mental model theory of extensional reasoning. *Psychological Review*, 106, 62–88.
- Johnson-Laird, P. N., & Savary, F. (1996). Illusory inferences about probabilities. *Acta Psychologica*, 93, 69–90.
- Johnson-Laird, P. N., & Savary, F. (1999). Illusory inferences: A novel class of erroneous deductions. *Cognition*, 71, 191–229.
- Johnson-Laird, P. N., & Stevenson, R. (1970). Memory for syntax. *Nature*, 227, 412.
- Johnson-Laird, P. N., & Wason, P. C. (Eds.). (1977). *Thinking*. Cambridge, England: Cambridge University Press.
- Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, England: Cambridge University Press.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237–251.
- Kamp, H. (1981). A theory of truth and semantic representation. In J. A. G. Groenendijk, T. M. V. Janssen, & M. B. J. Stokhof (Eds.), *Formal methods in the study of language* (pp. 277–322). Amsterdam, The Netherlands: Mathematical Centre Tracts.
- Karttunen, L. (1976). Discourse referents. In J. D. McCawley (Ed.), *Syntax and semantics, Vol. 7: Notes from the linguistic underground* (pp. 363–386). New York: Academic Press.
- Khemlani, S., & Johnson-Laird, P. N. (2009). Disjunctive illusory inferences and how to eliminate them. *Memory and Cognition*, 37, 615–623.
- Knauff, M., Fangmeier, T., Ruff, C. C. & Johnson-Laird, P. N. (2003). Reasoning, models, and images: Behavioral measures and cortical activity. *Journal of Cognitive Neuroscience*, 4, 559–573.
- Knauff, M., & Johnson-Laird, P. N. (2002). Imagery can impede inference. *Memory & Cognition*, 30, 363–371.
- Köhler, W. (1938). *The place of value in a world of facts*. New York: Liveright.
- Kroger, J. K., Nystrom, L. E., Cohen, J. D., & Johnson-Laird, P. N. (2008). Distinct neural substrates for deductive and mathematical processing. *Brain Research*, 1243, 86–103.
- Lagnado, D. A., Waldmann, M. R., Hagmayer Y., & Sloman, S. A. (2007). Beyond covariation: Cues to causal structure. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 154–172). Oxford, England: Oxford University Press.
- Landau, B., Spelke, E., & Gleitman, H. (1984). Spatial knowledge in a young blind child. *Cognition*, 16, 225–260.
- Lee, N. Y. L., & Johnson-Laird, P. N. (2006). Are there cross-cultural differences in reasoning? In *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*, 459–464.
- Legrenzi, P., Girotto, V., & Johnson-Laird, P. N. (2003). Models of consistency. *Psychological Science*, 14, 131–137.
- Mackiewicz, R., & Johnson-Laird, P. N. (2004). How order of information affects reasoning. *Polish Psychological Bulletin*, 35, 197–208.
- Markovits, H. (1984). Awareness of the “possible” as a mediator of formal thinking in conditional reasoning problems. *British Journal of Psychology*, 75, 367–376.
- Marr, D. (1982). *Vision*. San Francisco, CA: Freeman.
- Maxwell, J. C. (1911). Diagram. *The Encyclopaedia Britannica, Vol. XVII*. New York: The Encyclopaedia Britannica Company.
- McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naïve beliefs about the motions of objects. *Science*, 210, 1139–1141.
- Metzler, J., & Shepard, R.N. (1982). Transformational studies of the internal representations of three-dimensional objects. In R. N. Shepard & L. A. Cooper (Eds.), *Mental images and their transformations* (pp. 25–71). Cambridge, MA: MIT Press.
- Mill, J. S. (1874). *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific evidence*. (8th ed.). New York: Harper.
- Moray, N. (1990). A lattice theory approach to the structure of mental models. *Philosophical Transactions of the Royal Society of London B*, 327, 577–583.
- Moray, N. (1999). Mental models in theory and practice. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 223–258). Cambridge, MA: MIT Press.
- Morris, B. J., & Sloutsky, V. (2002). Children’s solutions of logical versus empirical problems: What’s missing and what develops? *Cognitive Development*, 16, 907–928.
- Neth, H., & Johnson-Laird, P. N. (1999). The search for counterexamples in human reasoning. In *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society*, 806.
- Newstead, S. E., Ellis, M. C., Evans, J. St. B. T., & Dennis, I. (1997). Conditional reasoning with realistic material. *Thinking & Reasoning*, 3, 49–76.
- Newstead, S. E., & Griggs, R. A. (1999). Premise misinterpretation and syllogistic reasoning. *Quarterly Journal of Experimental Psychology*, 52A, 1057–1075.
- Newstead, S. E., Handley, S. J., & Buck, E. (1999). Falsifying mental models: Testing the predictions of theories of syllogistic reasoning. *Memory & Cognition*, 27, 344–354.
- Oaksford, M., & Chater, N. (2007). *Bayesian rationality*. New York: Oxford University Press.
- Ormerod, T. C., Manktelow, K. I., & Jones, G. V. (1993). Reasoning with three types of conditional: Biases and mental models. *Quarterly Journal of Experimental Psychology*, 46A, 653–678.
- Osherson, D. N. (1974–1976). *Logical abilities in children* (Vols. 1–4). Hillsdale, NJ: Erlbaum.

- Peirce, C. S. (1931–1958). *Collected papers of Charles Sanders Peirce*. C. Hartshorne, P. Weiss, & A. Burks (Eds.). Cambridge, MA: Harvard University Press.
- Peirce, C. S. (1955). Abduction and induction. In J. Buchler (Ed.), *Philosophical writings of Peirce* (pp. 150–156). New York: Dover. [Original work published in 1903].
- Peng, K., & Nisbett, R. E. (1999). Culture, dialectics, and reasoning about contradiction. *American Psychologist*, 54, 741–754.
- Perales, J. C., & Shanks, D. R. (2008). Driven by power? Probe question and presentation format effects on causal judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34, 1482–1494.
- Polk, T. A., & Newell, A. (1995). Deduction as verbal reasoning. *Psychological Review*, 102, 533–566.
- Popper, K. R. (1972). *Objective knowledge*. Oxford, England: Clarendon.
- Quelhas, A.C., Johnson-Laird, P. N., & Juhasz, C. (2010). The modulation of conditional assertions and its effects on reasoning. *Quarterly Journal of Experimental Psychology*, 63, 1716–1739.
- Richardson, J., & Ormerod, T. C. (1997). Rephrasing between disjunctives and conditionals: Mental models and the effects of thematic content. *Quarterly Journal of Experimental Psychology*, 50A, 358–385.
- Roberts, M. J. (2000). Strategies in relational inference. *Thinking & Reasoning*, 6, 1–26.
- Roberts, M. J. (2003). Falsification and mental models: it depends on the task. In W. Schaeken, A. Vandierendonck, W. Schroyens, & G. d'Ydewalle (Eds.), *The mental models theory of reasoning: Refinement and extensions* (pp. 85–113). Mahwah, NJ: Erlbaum.
- Rouse, W.B., & Hunt, R.M. (1984). Human problem solving in fault diagnosis tasks. In W. B. Rouse (Ed.), *Advances in man-machine systems research* (pp. 195–222). Greenwich, CT: JAI Press.
- Santamaría, C., & Johnson-Laird, P. N. (2000). An antidote to illusory inferences. *Thinking & Reasoning*, 6, 313–333.
- Schaeken, W. S., De Vooght, G., Vandierendonck, A., & d'Ydewalle, G. (Eds.). (1999). *Deductive reasoning and strategies*. Mahwah, NJ: Erlbaum.
- Schaeken, W. S., Johnson-Laird, P. N., d'Ydewalle, G. (1996). Mental models and temporal reasoning. *Cognition*, 60, 205–234.
- Schroyens, W., & Schaeken, W. (2003). A critique of Oaksford, Chater, and Larkin's (2000) conditional probability model of conditional reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 140–149.
- Schroyens, W., Schaeken, W., & Dieussaert, K. (2008). "The" interpretation(s) of conditionals. *Experimental Psychology*, 58, 173–181.
- Schwartz, D., & Black, J. B. (1996). Analog imagery in mental model reasoning: Depictive models. *Cognitive Psychology*, 30, 154–219.
- Shimojo, S., & Ichikawa, S. (1989). Intuitive reasoning about probability: Theoretical and experimental analyses of the 'problem of three prisoners'. *Cognition*, 32, 1–24.
- Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119, 3–22.
- Sloutsky, V. M., & Goldvarg, Y. (1999). Effects of externalization on representation of indeterminate problems. In M. Hahn & S. Stones (Eds.), *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp. 695–700). Mahwah, NJ: Erlbaum.
- Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.
- Stenning, K., & Yule, P. (1997). Image and language in human reasoning: A syllogistic illustration. *Cognitive Psychology*, 34, 109–159.
- Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam, The Netherlands: North-Holland.
- Téglás, E., Girotto, V., Gonzalez, M., & Bonatti, L. L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proceedings of the National Academy of Sciences*, 104, 19156–19159.
- Tversky, A., & Kahneman, D. (1982). Causal schemas in judgments under uncertainty. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 117–128). Cambridge, England: Cambridge University Press.
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90, 292–315.
- Vadeboncoeur, I., & Markovits, H. (1999). The effect of instructions and information retrieval on accepting the premises in a conditional reasoning task. *Thinking & Reasoning*, 5, 97–113.
- Van der Henst, J.-B., Yang, Y., & Johnson-Laird, P. N. (2002). Strategies in sentential reasoning. *Cognitive Science*, 26, 425–468.
- Vandierendonck, A., & De Vooght, G. (1997). Working memory constraints on linear reasoning with spatial and temporal contents. *Quarterly Journal of Experimental Psychology*, 50A, 803–820.
- Vandierendonck, A., De Vooght, G., Desimpelaere, C., & Dierckx, V. (1999). Model construction and elaboration in spatial linear syllogisms. In W. S. Schaeken, G. De Vooght, A. Vandierendonck, & G. d'Ydewalle (Eds.), *Deductive reasoning and strategies* (pp. 191–207). Mahwah, NJ: Erlbaum.
- Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology*, 24, 535–585.
- Walsh, C., & Johnson-Laird, P. N. (2004). Co-reference and reasoning. *Memory & Cognition*, 32, 96–106.
- Wason, P. C., & Johnson-Laird, P. N. (1972). *The psychology of reasoning*. London: Batsford; Cambridge, MA: Harvard University Press.
- Webber, B. L. (1978). Description formation and discourse model synthesis. In D. L. Waltz (Ed.), *Theoretical issues in natural language processing*. New York: Association for Computing Machinery.
- Wetherick, N. E., & Gilhooly, K. J. (1995). "Atmosphere", matching, and logic in syllogistic reasoning. *Current Psychology: Developmental, Learning, Personality, Social*, 14, 169–178.
- Wittgenstein, L. (1922). *Tractatus logico-philosophicus*. London: Routledge & Kegan Paul.
- Yang, Y., & Johnson-Laird, P. N., (2000). How to eliminate illusions in quantified reasoning, *Memory & Cognition*, 28, 1050–1059.

Robert L. Goldstone and Ji Yun Son

Abstract

Humans and other animals perceive and act on the basis of similarities among things because similarities are usually informative. Similar things usually behave similarly, and because we can grasp these similarities, we can organize and predict the things in our world. Four major classes of models have been proposed for how humans assess similarities. In geometric models, entities are represented by their positions in a multidimensional space, and similarity is based on the proximity of entities in this space. In featural models, entities are described by their features, and the similarity of entities is an increasing function of their shared features and/or a decreasing function of their unique features. In alignment-based models, the similarity between two structured entities is calculated by placing the elements of their structures into correspondence. In transformational models, the similarity between two entities is conceptualized as the number of transformations required to transform one entity into the other. We discuss issues related to the flexibility and constraints of similarity, and how similarity grounds other cognitive processes.

Key Words: similarity, multidimensional scaling, induction, categorization, contrast model, relations, alignment, structural representations, transformations, grounding

Introduction

Human assessments of similarity are fundamental to cognition because similarities in the world are revealing. The world is an orderly enough place that similar objects and events tend to behave similarly. This fact of the world is not just a fortunate coincidence. It is *because* objects are similar that they will tend to behave similarly in most respects. It is because crocodiles and alligators are similar in their external form, internal biology, behavior, diet, and customary environment that one can often successfully generalize from what one knows of one to the other. As Quine (1969) observed, "Similarity is fundamental for learning, knowledge and thought, for only our sense of similarity allows us to order things into kinds so that these can function as stimulus meanings. Reasonable expectation depends on the similarity of circumstances and on our tendency to

expect that similar causes will have similar effects (p. 114)." Similarity thus plays a crucial role in making predictions because similar things usually behave similarly.

From this perspective, psychological assessments of similarity are valuable to the extent that they provide grounds for predicting as many important aspects of our world as possible (Griffiths et al., Chapter 3; Holland, Holyoak, Nisbett, & Thagard, 1986; Dunbar & Klahr, Chapter 35). Appreciating the similarity between crocodiles and alligators is helpful because information learned about one is generally true of the other. If we learned an arbitrary fact about crocodiles, such as they are very sensitive to the cold, then we would probably be safe in inferring that this fact is also true of alligators. As the similarity between A and B increases, so does the probability of correctly inferring that B has X upon knowing

that A has X (Tenenbaum, 1999). This relation assumes that we have no special knowledge related to property X. Empirically, Heit and Rubinstein (1994) have shown that if we *do* know about the property, then this knowledge, rather than a one-size-fits-all similarity, is used to guide our inferences. For example, if people are asked to make an inference about an anatomical property, then anatomical similarities have more influence than behavioral similarities. Boars are anatomically but not behaviorally similar to pigs, and this difference successfully predicts that people are likely to make anatomical but not behavioral inferences from pigs to boars. The logical extreme of this line of reasoning (Goodman, 1972; Quine, 1977) is that if one has complete knowledge about the reasons why an object has a property, then general similarity is no longer relevant to generalizations. The knowledge itself completely guides whether the generalization is appropriate. Moonbeams and melons are not very similar generally speaking, but if one is told that the Moonbeams have the property that the word begins with Melanie's favorite letter, then one can generalize this property to melons with very high confidence.

By contrasting the cases of crocodiles, boars, and moonbeams, we can specify the benefits and limitations of similarity. We tend to rely on similarity to generate inferences and categorize objects into kinds when we do not know exactly what properties are relevant, we do not have explicit causal accounts for what makes something appear the way it does (Buehner & Cheng, Chapter 12), or when we cannot easily separate an object into separate properties. Similarity is an excellent example of a domain-general source of information. Even when we do not have specific knowledge about a domain, we can use similarity as a default method to reason about it. The contravening limitation of this domain generality is that when specific knowledge is available, then a generic assessment of similarity is no longer as relevant (Keil, 1989; Lombrozo, Chapter 14; Murphy, 2002; Murphy & Medin, 1985; Rips, 1989; Rips & Collins, 1993). Artificial laboratory experiments where subjects are asked to categorize unfamiliar stimuli into novel categories invented by the experimenter are situations where similarity is clearly important because subjects have little else to use (Estes, 1994; Nosofsky, 1984, 1986; see Rips et al., Chapter 11). However, similarity is also important in many real-world situations because our knowledge does not run as deep as we think it does (Rozenblit & Keil,

2002), and because a general sense of similarity often has an influence even when more specific knowledge ought to overrule it (Allen & Brooks, 1991; Smith & Sloman, 1994).

Another argument for the importance of similarity in cognition is simply that it plays a significant role in psychological accounts of problem solving, memory, prediction, and categorization. If a problem is similar to a previously solved problem, then the solution to the old problem may be applied to the new problem (Holyoak & Koh, 1987; Ross, 1987, 1989; see Holyoak, Chapter 13). If a cue is similar enough to a stored memory, the memory may be retrieved (Raaijmakers & Shiffrin, 1981). If an event is similar enough to a previously experienced event, the stored event's outcome may be offered as a candidate prediction for the current event (Sloman, 1993; Tenenbaum & Griffiths, 2001). If an unknown object is similar enough to a known object, then the known object's category label may be applied to the unknown object (Nosofsky, 1986). The act of comparing events, objects, and scenes, and establishing similarities between them is of critical importance for the cognitive processes we depend upon.

The utility of similarity for grounding our concepts has been rediscovered in all of the fields comprising cognitive science (see Medin & Rips, Chapter 2). Exemplar (Estes, 1994; Kruschke, 1992; Lamberts, 2000; Medin & Schaffer, 1978; Nosofsky, 1986), instance-based (Aha, 1992), view-based (Tarr & Gauthier, 1998), case-based (Schank, 1982), nearest neighbor (Ripley, 1996), configural cue (Gluck & Bower, 1990), and vector quantization (Kohonen, 1995) models all share the underlying strategy of giving responses learned from similar, previously presented patterns to novel patterns. Thus, a model can respond to repetitions of these patterns; it can also give responses to novel patterns that are likely to be correct by sampling responses to old patterns, weighted by their similarity to the novel pattern. Consistent with these models, psychological evidence suggests that people show good transfer to new stimuli in perceptual tasks to the extent that the new stimuli resemble previously learned stimuli (Kolers & Roediger, 1984; Palmeri, 1997). Another common feature of these approaches is that they represent patterns in a relatively raw, unprocessed form. This parallels the constraint described earlier on the applicability of similarity. Both raw representations and generic similarity assessments are most useful as a default strategy when one does not know

exactly what properties of a stimulus are important. One's best bet is to follow the principle of least commitment (Marr, 1982) and keep mental descriptions in a relatively raw form to preserve information that may be needed at a later point.

Another reason for studying similarity is that it provides an elegant diagnostic tool for examining the structure of our mental entities and the processes that operate on them. For example, one way to tell that a physicist has progressed beyond the novice stage is that he or she sees deep similarities between problems that require calculation of force even though the problems are superficially dissimilar (Chi, Feltovich, & Glaser, 1981; see Bassok & Novick, Chapter 21). Given that psychologists have no microscope with direct access to people's representations of their knowledge, appraisals of similarity provide a powerful, if indirect, lens onto representation/process assemblies (see also Markman, Chapter 4; Doumas & Hummel, Chapter 5).

A final reason to study similarity is that it occupies an important ground between perceptual constraints and the functions of higher level knowledge systems. Similarity is grounded by perceptual functions. A tone of 200 Hz and a tone of 202 Hz sound similar (Shepard, 1987), and the similarity is cognitively impenetrable (Pylyshyn, 1985) enough that there is little that can be done to alter this perceived similarity. On the other hand, similarity is also highly flexible and dependent on knowledge and purpose. By focusing on patterns of motion and relations, even electrons and planets can be made to seem similar (Gentner, 1983; Holyoak & Thagard, 1989; see Holyoak, Chapter 13). A complete account of similarity will make contact both with Fodor's (1983) isolated and modularized perceptual input devices and the "central system" where everything a person knows may be relevant.

A Survey of Major Approaches to Similarity

There have been a number of formal treatments that simultaneously provide theoretical accounts of similarity and describe how it can be empirically measured (Hahn, 2003). These models have had a profound practical impact in statistics, automatic pattern recognition by machines, data mining, and marketing (e.g., online stores can inform you that "people similar to you liked the following other items..."). Our brief survey is organized in terms of the following models: geometric, feature-based, alignment-based, and transformational.

Geometric Models and Multidimensional Scaling

Geometric models of similarity have been among the most influential approaches to analyzing similarity (Carroll & Wish, 1974; Torgerson, 1958, 1965). These approaches are exemplified by nonmetric multidimensional scaling (MDS) models (Shepard, 1962a, 1962b). MDS models represent similarity relations between entities in terms of a geometric model that consists of a set of points embedded in a dimensionally organized metric space. The input to MDS routines may be similarity judgments, dissimilarity judgments, confusion matrices, correlation coefficients, joint probabilities, or any other measure of pairwise proximity. The output of an MDS routine is a geometric model of the data, with each object of the data set represented as a point in an n -dimensional space. The similarity between a pair of objects is taken to be inversely related to the distance between two objects' points in the space. In MDS, the distance between points i and j is typically computed by:

$$distance(i, j) = \left[\sum_{k=1}^n |X_{ik} - X_{jk}|^r \right]^{\frac{1}{r}} \quad (\text{Eq. 1})$$

where n is the number of dimensions, X_{ik} is the value of dimension k for item i , and r is a parameter that allows different spatial metrics to be used. With $r = 2$, a standard Euclidean notion of distance is invoked, whereby the distance between two points is the length of the straight line connecting the points. If $r = 1$, then distance involves a city-block metric where the distance between two points is the sum of their distances on each dimension ("short-cut" diagonal paths are not allowed to directly connect points differing on more than one dimension). A Euclidean metric often provides a better fit to empirical data when the stimuli being compared are composed of integral, perceptually fused dimensions such as the brightness and saturation of a color. Conversely, a city-block metric is often appropriate for psychologically separated dimensions such as brightness and size (Attnave, 1950).

Richardson's (1938) fundamental insight, which is the basis of contemporary use of MDS, was to begin with subjects' judgments of pairwise object dissimilarity and work backward to determine the dimensions and dimension values that subjects used in making their judgments. MDS algorithms proceed by placing entities in an N -dimensional space such that the distances between the entities

accurately reflect the empirically observed similarities. For example, if we asked people to rate the similarities [on a scale from 1 (low similarity) to 10 (high similarity)] of Russia, Cuba, and Jamaica, we might find:

$$\begin{aligned}\text{Similarity (Russia, Cuba)} &= 7 \\ \text{Similarity (Russia, Jamaica)} &= 1 \\ \text{Similarity (Cuba, Jamaica)} &= 8\end{aligned}$$

An MDS algorithm would work to position the three countries in a space such that countries that are rated as being highly similar are very close to each other in the space. With nonmetric scaling techniques only ordinal similarity relations are preserved. The interpoint distances suggested by the similarity ratings may not be simultaneously satisfiable in a given dimensional space. If we limit ourselves to a single dimension (we place the countries on a “number line”), then we cannot simultaneously place Russia near Cuba (similarity = 7) and place Russia far away from Jamaica (similarity = 1). In MDS terms, the “stress” of the one-dimensional solution would be high. We could increase the dimensionality of our solution and position the points in two-dimensional space. A perfect reconstruction of any set of proximities among a set of N objects can be obtained if a high enough dimensionality (specifically, $N - 1$ dimensions) is used. However, increasing the number of dimensions always risks explaining noise in the data (“overfitting”) and tends to yield less interpretable solutions.

One of the main applications of MDS is to determine the underlying dimensions comprising the set of compared objects. Once the points are positioned in a way that faithfully mirrors the subjectively obtained similarities, it is often possible to give interpretations to the axes or to rotations of the axes. In the earlier example, dimensions may correspond to “political affiliation” and “climate.” Russia and Cuba would have similar values on the former dimension; Jamaica and Cuba would have similar values on the latter dimension. A study by Smith, Shoben, and Rips (1974) illustrates a classic use of MDS. They obtained similarity ratings from subjects on many pairs of birds. Submitting these pairwise similarity ratings to MDS analysis, they hypothesized underlying features that were used for representing the birds. Assigning subjective interpretations to the geometric model’s axes, the experimenters suggested that birds were represented in terms of their values on dimensions such as “ferocity” and “size.”

As another example of using MDS to create quantitative object and dimension interpretations, Figure 10.1 presents an MDS solution for 135 words representing animals. We started by obtaining subjective estimates of similarity by randomly selecting six anchor animal terms and asking subjects to determine to which of these terms each of the remaining 129 animals terms was most similar. We considered two terms as belonging to the same category if the same anchor term was judged most similar to both of them. After several passes through the 135 animals, choosing different random anchor terms each time, we calculated the similarity between any pair of animals as their probability of being placed in the same category. The resulting MDS solution based on this $(135 \times 135)/2$ (divided by 2 because we assume the similarity of A to B is equal to the similarity of B to A) matrix of similarities is shown in Figure 10.1. The first four dimensions can be imperfectly interpreted as mammalian versus nonmammalian (Dimension 1, with mammals on the left and nonmammals on the right), water-land-air (Dimension 2, with flying animals at the top and water animals at the bottom), mundane-mythical (Dimension 3, with mythical or extinct animals on the left), and unpleasant-pleasant (Dimension 4, with unpleasant animals at the bottom).

It is important to note that the proper psychological interpretation of a geometric representation of objects is not necessarily in terms of its Cartesian axes. In some domains, such as musical pitches, the best interpretation of objects may be in terms of their polar coordinates of angle and length (Shepard, 1982). Recent work has extended geometric representations still further, representing patterns of similarities by generalized, nonlinear manifolds (Tenenbaum, De Silva, & Lanford, 2000).

MDS is also used to create a compressed representation that conveys relational similarities among a set of items. A set of N items requires $N(N-1)/2$ numbers to express all pairwise distances among the items, assuming that any object has a distance of 0 to itself and distances are symmetric. However, if an MDS solution fits the distance data well, it can allow these same distances to be reconstructed using only ND numbers, where D is the number of dimensions of the MDS solution. This compression may be psychologically very useful. One of the main goals of psychological representation is to create efficient codes for representing a set of objects. Compressed representations can facilitate encoding, memory, and processing. Shimon Edelman (1999)

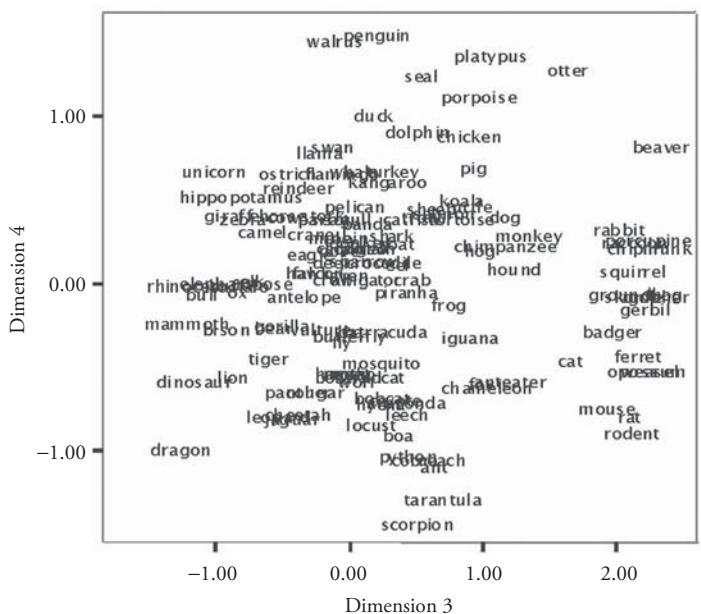
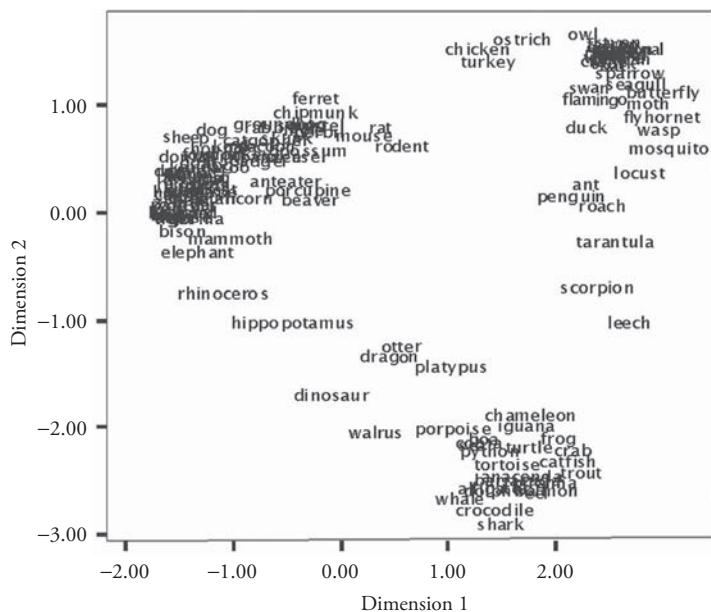


Fig. 10.1 A four-dimensional multidimensional scaling (MDS) solution for 135 animal terms. The four dimensions, in order, may be imperfectly interpreted as reflecting mammalian versus nonmammalian, water-land-air, mundane-mythical, and unpleasant-pleasant.

has proposed that both people and machines efficiently code their world by creating geometric spaces for objects with much lower dimensionality than the objects' physical description (see also Gardenfors, 2000).

A third use of MDS is to create quantitative representations that can be used in mathematical and computational models of cognitive processes. Numeric representations, namely coordinates in a psychological space, can be derived for stories, pictures, sounds, words, or any other stimuli for which

one can obtain subjective similarity data. Once constructed, these numeric representations can be used to predict people's categorization accuracy, memory performance, or learning speed. MDS models have been successful in expressing cognitive structures in stimulus domains as far removed as animals (Smith et al., 1974), Rorschach ink blots (Osterholm, Woods, & Le Unes, 1985), chess positions (Horgan, Millis, & Neimeyer, 1989), and air flight scenarios (Schvaneveldt, 1985). Many objects, situations, and concepts seem to be psychologically structured in

terms of dimensions, and a geometric interpretation of the dimensional organization captures a substantial amount of that structure.

Featural Models

In 1977, Amos Tversky brought into prominence what would become the main contender to geometric models of similarity in psychology. The reason given for proposing a feature-based model was that subjective assessments of similarity did not always satisfy the assumptions of geometric models of similarity.

PROBLEMS WITH THE STANDARD GEOMETRIC MODEL

Three assumptions of standard geometric models of similarity are as follows:

Minimality: $D(A,B) \geq D(A,A) = 0$

Symmetry: $D(A,B) = D(B,A)$

The Triangle Inequality: $D(A,B) + D(B,C) \geq D(A,C)$

where $D(A,B)$ is interpreted as the dissimilarity between items A and B.

According to the minimality assumption, all objects are equally (dis)similar to themselves. Some violations of this assumption are found (Nickerson, 1972) when confusion rates or RT measures of similarity are used. First, not all letters are equally similar to themselves. For example, Podgorny and Garner (1979) found that if the letter S is shown twice on a screen, subjects are faster to correctly say that the two tokens are similar (i.e., they come from the same similarity-defined cluster) than if the twice-shown letter is W. By this reaction time measure of similarity, the letter S is more similar to itself than the letter W is to itself. Even more troublesome for the minimality assumption, two different letters may be more similar to each other than a particular letter is to itself. For example, the letter C is more similar to the letter O than W is to itself, as measured by inter-letter confusions. Gilmore, Hersh, Caramazza, and Griffin (1979) found that the letter M is more often recognized as an H (probability = .391) than as an M (probability = .180). This finding is problematic for geometric representations because the distance between a point and itself should be zero.

According to the symmetry assumption, (dis)similarity should not be affected by the ordering of items because the distance from Point A to B is equal to the distance from B to A. Contrary to this presumed symmetry, similarity is asymmetric on occasion (Tversky, 1977). In one of Tversky's

examples, North Korea is judged to be more similar to Red China than Red China is to North Korea. Often, a nonprominent item is more similar to a prominent item than vice versa. This is consistent with the result that people judge their friends to be more similar to themselves than they themselves are to their friends (Holyoak & Gordon, 1983), under the assumption that a person is highly prominent to him/herself. Polk et al. (2002) found that when the frequency of colors was experimentally manipulated, rare colors are judged to be more similar to common colors than common colors are to rare colors.

According to the triangle inequality assumption, the distance/dissimilarity between two points A and B cannot be more than the distance between A and a third point C plus the distance between C and B. Geometrically speaking, a straight line connecting two points is the shortest path between the points. Tversky and Gati (1982) find violations of this assumption when it is combined with an assumption of segmental additivity ($D(A,B) + D(B,C) = D(A,C)$) if A, B, and C lie on a straight line). Consider three items in multidimensional space, A, B, and C, falling on a straight line such that B is between A and C. Also consider a fourth point, E, that forms a right triangle when combined with A and C. The triangle inequality assumption *cum* segmental additivity predicts that

$$D(A,E) \geq D(A,B) \text{ and } D(E,C) \geq D(B,C)$$

or

$$D(A,E) \geq D(B,C) \text{ and } D(E,C) \geq D(A,B)$$

Systematic violations of this prediction are found such that the path going through the corner point E is shorter than the path going through the center point B. For example, if the items represented by their coordinates in the size-color of Figure 10.2 are instantiated as follows:

A= White, three inches

B= Pink, four inches

C= Red, five inches

E= Red, three inches

then people's dissimilarity ratings have empirically demonstrated violations of the triangle inequality such that $D(A,E) < D(A,B)$ and $D(E,C) < D(B,C)$ (Tversky & Gati, 1982). Such an effect can be modeled by geometric models of similarity if r in Eq. 1 is given a value less than 1. However, if r is less than one, then dissimilarity does not satisfy a power metric, often times considered a minimal assumption

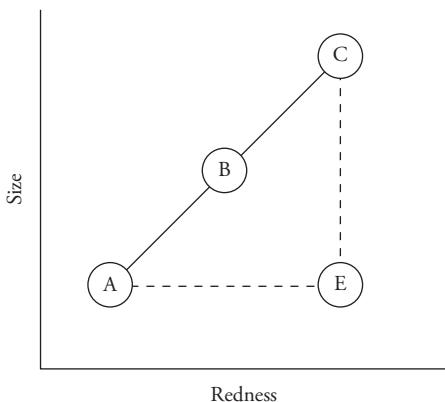


Fig. 10.2 The triangle inequality assumption requires the path from A to C going through B to be shorter than or equal to the path from A to C going through E.

for geometric solutions to be interpretable. The two assumptions of a power metric are (1) distances along straight lines are additive, and (2) the shortest path between points is a straight line.

Other potential problems with geometric models of similarity are (1) they strictly limit the number of nearest neighbors an item can have (Tversky & Hutchinson, 1986), (2) multidimensional scaling techniques have difficulty describing items that vary on a large number of features (Krumhansl, 1978), and (3) standard MDS techniques do not predict that adding common features to items increases their similarity (Tversky & Gati, 1982). On the first point: MDS models consisting of two dimensions cannot predict that item X is the closest item to 100 other items. There would be no way of placing those 100 items in two dimensions such that X is closer to all of them than any other item is. For human data, a superordinate term (e.g., fruit) is often the nearest neighbor of many of its exemplars (apples, bananas, etc.) as measured by similarity ratings. On the second point: Although there is no logical reason why geometric models cannot represent items of any number of dimensions (as long as the number of dimensions is less than number of items minus one), geometric models tend to yield the most satisfactory and interpretable solutions in low-dimensional space. MDS solutions involving more than six dimensions are rare. On the third point: The addition of the same feature to a pair of items increases their rated similarity (Gati & Tversky, 1984), but this is incompatible with simple MDS models. If adding a shared feature corresponds to adding a dimension in which the two items under consideration have the same value, then

there will be no change to the items' dissimilarity because the geometric distance between the points remains the same. MDS models that incorporate the dimensionality of the space could predict the influence of shared features on similarity, but such models would no longer relate similarity directly to an inverse function of interitem distance.

One research strategy has been to augment geometrical models of similarity in ways that solve these problems. One solution, suggested by Carol Krumhansl (1978), has been to model dissimilarity in terms of both interitem distance in a multidimensional space *and* spatial density in the neighborhoods of the compared items. The more items there are in the vicinity of an item, the greater is the spatial density of the item. Items are more dissimilar if they have many items surrounding them (their spatial density is high) than if they have few neighboring items. By including spatial density in an MDS analysis, violations of minimality, symmetry, and the triangle inequality can potentially be accounted for, as well as some of the influence of context on similarity. However, the empirical validity of the spatial density hypothesis is in some doubt (Corter, 1987, 1988; Krumhansl, 1988; Tversky & Gati, 1982).

Robert Nosofsky (1991) has suggested another potential way to save MDS models from some of the aforementioned criticisms. He introduces individual bias parameters in addition to the interitem relation term. Similarity is modeled in terms of interitem distance *and* in terms of biases toward particular items. Biases toward items may be due to attention, salience, knowledge, and frequency of items. This revision handles asymmetric similarity results and the result that a single item may be the most similar item to many other items, but it does not directly address several of the other objections.

THE CONTRAST MODEL

In light of the aforementioned potential problems for geometric representations, Tversky (1977) proposed to characterize similarity in terms of a feature-matching process based on weighting common and distinctive features. In this model, entities are represented as a collection of features and similarity is computed by

$$S(A,B) = \theta f(A \cap B) - \alpha f(A-B) - \beta f(B-A).$$

The similarity of A to B is expressed as a linear combination of the measure of the common and

distinctive features. The term $(A \cap B)$ represents the features that items A and B have in common. $(A - B)$ represents the features that A has but B does not. $(B - A)$ represents the features of B that are not in A. θ , α , and β are weights for the common and distinctive components. Common features, as compared to distinctive features, are given relatively more weight for verbal as opposed to pictorial stimuli (Gati & Tversky, 1984), for cohesive as opposed to noncohesive stimuli (Ritov, Gati, & Tversky, 1990), for similarity as opposed to difference judgments (Tversky, 1977), and for entities with a large number of distinctive as opposed to common features (Gati & Tversky, 1984). There are no restrictions on what may constitute a feature. A feature may be any property, characteristic, or aspect of a stimulus. Features may be concrete or abstract (i.e., "symmetric" or "beautiful").

The Contrast Model predicts asymmetric similarity because A is not constrained to equal B and $f(A - B)$ may not equal $f(B - A)$. North Korea is predicted to be more similar to Red China than vice versa if Red China has more salient distinctive features than North Korea, and A is greater than B. The Contrast Model can also account for nonmirroring between similarity and difference judgments. The common features term $(A \cap B)$ is hypothesized to receive more weight in similarity than difference judgments; the distinctive features terms receive relatively more weight in difference judgments. As

a result, certain pairs of stimuli may be perceived as simultaneously being more similar to and more different from each other, compared to other pairs (Tversky, 1977). Sixty-seven percent of a group of subjects selected West Germany and East Germany as more similar to each other than are Ceylon and Nepal. Seventy percent of subjects also selected West Germany and East Germany as more different from each other than are Ceylon and Nepal. According to Tversky, East and West Germany have more common and more distinctive features than Ceylon and Nepal. Medin, Goldstone, and Gentner (1993) present additional evidence for nonmirroring between similarity and difference, exemplified in Figure 10.3. When two scenes share a relatively large number of relational commonalities (e.g., Scenes T and B both have three objects that have the *same pattern*) but also a large number of differences on specific attributes (e.g., none of the patterns in Scene T match any of the patterns in B), then the scenes tend to be judged as simultaneously very similar and very different.

A number of models are similar to the Contrast model in basing similarity on features and in using some combination of the $(A \cap B)$, $(A - B)$, and $(B - A)$ components. Sjoberg (1972) proposes that similarity is defined as $f(A \cap B)/f(A \cup B)$. Eisler and Ekman (1959) claim that similarity is proportional to $f(A \cap B)/(f(A) + f(B))$. Bush and Mosteller (1951) defines similarity as $f(A \cap B)/f(A)$. These three models

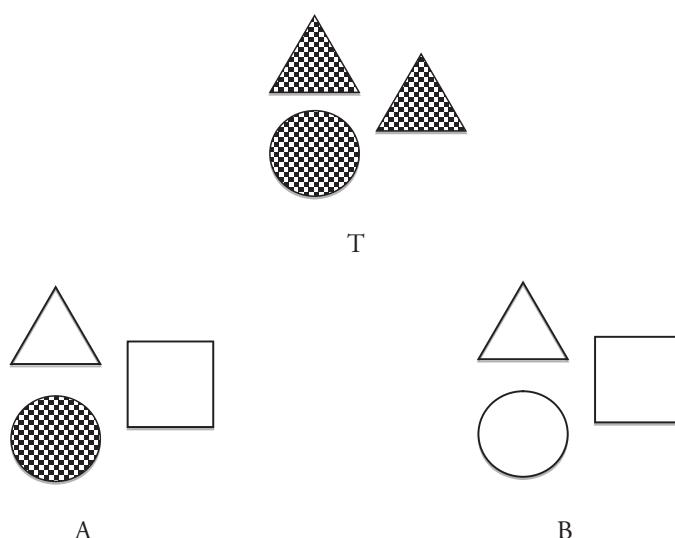


Fig. 10.3 The set of objects in B is selected as both more similar to, and more different from, the set of objects in T.

can all be considered specializations of the general equation $f(A \cap B) / [f(A \cap B) + \alpha f(A - B) + \beta f(B - A)]$. As such, they differ from the Contrast model by applying a ratio function as opposed to a linear contrast of common and distinctive features.

The fundamental premise of the Contrast Model, that entities can be described in terms of constituent features, is a powerful idea in cognitive psychology. Featural analyses have proliferated in domains of speech perception (Jakobson, Fant, & Halle, 1963), pattern recognition (Neisser, 1967; Treisman, 1986), perception physiology (Hubel & Wiesel, 1968), semantic content (Katz & Fodor, 1963), and categorization (Medin & Shaffer, 1978). Neural network representations are often based on features, with entities being broken down into a vector of ones and zeros, where each bit refers to a feature or “micro-feature.” Similarity plays a crucial role in many connectionist theories of generalization, concept formation, and learning. The notion of dissimilarity used in these systems is typically the fairly simple function “Hamming distance.” The Hamming distance between two strings is simply their city-block distance; that is, it is their $(A - B) + (B - A)$ term. “1 0 0 1 1” and “1 1 1 1 1” would have a Hamming distance of 2 because they differ on two bits. Occasionally, more sophisticated measures of similarity in neural networks normalize dissimilarities by string length. Normalized Hamming distance functions can be expressed by $[(A - B) + (B - A)] / [f(A \cap B)]$.

Considerable recent progress on featural models of similarity has involved the development of mathematical and computational efforts to automatically construct the features that, if posited, could account for the similarity relations observed between a large set of objects (Lee, 1998). For example, Bayesian methods have been proposed that can determine both the number of features (Navarro & Griffiths, 2008) or dimensions (Lee, 2001) underlying a set of objects, and their importance for determining similarity. One principle derived from Bayesian models of similarity (see Griffiths et al., Chapter 3) that has received empirical support is the “size principle,” according to which rare features, when shared by two objects, increase their similarity more than do common features (Navarro & Perfors, 2010).

SIMILARITIES BETWEEN GEOMETRIC AND FEATURE-BASED MODELS

While MDS and featural models are often analyzed in terms of their differences, they also share

a number of similarities. Recent progress has been made on combining both representations into a single model, using Bayesian statistics to determine whether a given source of variation is more efficiently represented as a feature or dimension (Navarro & Lee, 2003). Tversky and Gati (1982) described methods of translating continuous dimensions into featural representations. Dimensions that are sensibly described as being more or less (e.g., loud is more sound than soft, bright is more light than dim, and large is more size than small) can be represented by sequences of nested feature sets. That is, the features of B include a subset of A’s features whenever B is louder, brighter, or larger than A. Alternatively, for qualitative attributes like shape or hue (red is not subjectively “more” than blue), dimensions can be represented by chains of features such that if B is between A and C on the dimension, then $(A \cap B) \supset (A \cap C)$ and $(B \cap C) \supset (A \cap C)$. For example, if orange lies between red and yellow on the hue dimension, then this can be featurally represented if orange and red share features that orange and yellow do not share.

An important attribute of MDS models is that they create *postulated* representations, namely dimensions that explain the systematicities present in a set of similarity data. This is a classic use of abductive reasoning: Dimensional representations are hypothesized that, if they were to exist, would give rise to the obtained similarity data. Other computational techniques share with MDS the goal of discovering the underlying descriptions for items of interest but create featural rather than dimensional representations. Hierarchical cluster analysis, like MDS, takes pairwise proximity data as input. Rather than output a geometric space with objects as points, hierarchical cluster analysis outputs an inverted-tree diagram, with items at the root-level connected with branches. The smaller the branching distance between two items, the more similar they are. Just as the dimensional axes of MDS solutions are given subjective interpretations, the branches are also given interpretations. For example, in Shepard’s (1972) analysis of speech sounds, one branch is interpreted as voiced phonemes while another branch contains the unvoiced phonemes. In additive cluster analysis (Shepard & Arabie, 1979) similarity data is transformed into a set of overlapping item clusters. Items that are highly similar will tend to belong to the same clusters. Each cluster can be considered as a feature. Recent progress has been made on efficient and mathematically principled models that

find such featural representations for large databases (Lee, 2002a, 2002b; Tenenbaum, 1996).

One particularly useful application of abducing representations that can account for patterns of similarity has been the development of automated techniques for analyzing large corpora of text. A computational approach to word meaning that has received considerable recent attention has been to base word meanings solely on the patterns of co-occurrence between a large number of words in an extremely large text corpus (Burgess & Lund, 2000; Griffiths, Steyvers, & Tenenbaum, 2007; Landauer & Dumais, 1997). Mathematical techniques are used to create vector encodings of words that efficiently capture their co-occurrences. If two words, such as “cocoon” and “butterfly,” frequently co-occur in an encyclopedia or enter into similar patterns of co-occurrence with other words, then their vector representations will be highly similar. The meaning of a word, its vector in a high dimensional space, is completely based on the contextual similarity of words to other words. Within this high dimensional space, Landauer and Dumais (1997) conceive of similarity as the cosine of the angle between two words rather than their distance. With these new techniques, it is now possible to create geometric spaces or featural encodings with tens of thousands of words.

Another commonality between geometric and featural representations, one that motivates the next major class of similarity models that we consider, is that both use relatively unstructured representations. Entities are structured as sets of features or dimensions with no relations between these attributes. Entities such as stories, sentences, natural objects, words, scientific theories, landscapes, and faces are not simply a “grab bag” of attributes. Two

kinds of structure seem particularly important: propositional and hierarchical. A proposition is an assertion about the relation between informational entities (Palmer, 1975; see Markman, Chapter 4; Doumas & Hummel, Chapter 5). For example, relations in a visual domain might include *Above*, *Near*, *Right*, *Inside*, and *Larger-than* that take informational entities as arguments. The informational entities might include features such as square and values on dimensions such as 3 inches. Propositions are defined as the smallest unit of knowledge that can stand as a separate assertion and have a truth value. The order of the arguments in the predicate is critical. For example, *Above* (*Triangle*, *Circle*) does not represent the same fact as *Above* (*Circle*, *Triangle*). Hierarchical representations involve entities that are embedded in one another. Hierarchical representations are required to represent the fact that X is *part of* Y or that X is a *kind of* Y. For example, in Collins and Quillian’s (1969) propositional networks, labeled links (“Is-a” links) stand for the hierarchical relation between *Canary* and *Bird*.

Some quick fixes to geometric and featural accounts of similarity are possible, but they fall short of a truly general capacity to handle structured inputs. Hierarchical clustering does create trees of features, but there is no guarantee that there are relationships, such as Is-a or Part-of, between the subtrees. However, structure might exist in terms of features that represent conjunctions of properties. For example, using the materials in Figure 10.4, 20 undergraduates were shown triads consisting of A, B, and T, and were asked to say whether Scene A or B was more similar to T. The strong tendency to choose A over B in the first panel suggests that the feature “square” influences similarity. Other

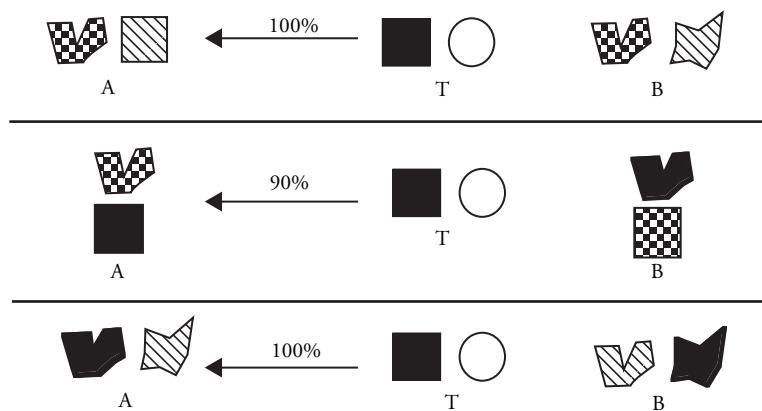


Fig. 10.4 The sets of objects T are typically judged to be more similar to the objects in the A sets than the B sets. These judgments show that people pay attention to more than just simple properties like “black” or “square” when comparing scenes.

choices indicated that subjects also based similarity judgments on the spatial locations and shadings of objects as well as their shapes.

However, it is not sufficient to represent the left-most object of T as {Left, Square, Black} and base similarity on the number of shared and distinctive features. In the second panel, A is again judged to be more similar to T than is B. Both A and B have the features “Black” and “Square.” The only difference is that for A and T, but not B, the “Black” and “Square” features belong to the same object. This judgment is only compatible with feature set representations if we include the possibility of *conjunctive features* in addition to *simple features* such as “Black” and “Square” (Gluck, 1991; Hayes-Roth & Hayes-Roth, 1977). By including the conjunctive feature “Black-Square,” possessed by both T and A, we can explain, using feature sets, why T is more similar to A than B. The third panel demonstrates the need for a “Black-Left” feature, and other data indicate a need for a “Square-Left” feature. Altogether, if we wish to explain the similarity judgments that people make, we need a feature set representation that includes six features (three simple and three complex) to represent the square of T.

However, there are two objects in T, bringing the total number of features required to at least two times the six features required for one object. The number of features required increases still further if we include feature-triplets such as “Left-Black-Square.” In general, if there are O objects in a scene, and each object has F features, then there will be OF simple features. There will be O conjunctive features that combine two simple features (i.e., *pairwise* conjunctive features). If we limit ourselves to simple and pairwise features to explain the pattern of similarity judgments in Figure 10.4, we still will require $OF(F+1)/2$ features per scene, or $OF(F+1)$ features for two scenes that are compared to one another.

Thus, featural approaches to similarity require a fairly large number of features to represent scenes that are organized into parts. Similar problems exist for dimensional accounts of similarity. The situation for these models becomes much worse when we consider that similarity is also influenced by relations between features, such as “Black to the left of white” and “square to the left of white.” Considering only binary relations, there are O^2F^2R -*OFF* relations within a scene that contains O objects, F features per object, and R different types of relations between features. Although more sophisticated objections have been raised about these approaches

by John Hummel and colleagues (Holyoak & Hummel, 2000; Hummel, 2000, 2001; Hummel & Biederman, 1992; Hummel & Holyoak, 1997, 2003; see Doumas & Hummel, Chapter 5), at the very least, geometric and featural models apparently require an implausibly large number of attributes to account for the similarity relations between structured, multipart scenes.

Alignment-Based Models

Partly in response to the difficulties that the previous models have in dealing with structured descriptions, a number of researchers have developed alignment-based models of similarity. In these models, comparison is not just matching features, but determining how elements correspond to, or align with, one another. Matching features are aligned to the extent that they play similar roles within their entities. For example, a car with a green wheel and a truck with a green hood both share the feature *green*, but this matching feature may not increase their similarity much because the car’s wheel does not correspond to the truck’s hood. Drawing inspiration from work on analogical reasoning (Gentner, 1983; Holyoak & Thagard, 1995; see also Holyoak, Chapter 13), in alignment-based models, matching features influence similarity more if they belong to parts that are placed in correspondence, and parts tend to be placed in correspondence if they have many features in common and are consistent with other emerging correspondences (Goldstone, 1994a; Markman & Gentner, 1993a). Alignment-based models make purely relational similarity possible (Falkenhainer, Forbus, & Gentner, 1989).

Initial evidence that similarity involves aligning scene descriptions was provided by Markman and Gentner (1993a), who found that when subjects are asked to determine corresponding objects, they tend to make more structurally sound choices when they have first judged the similarity of the scenes that contain the objects. For example, in Figure 10.5, subjects could be asked which object in the bottom set corresponds to the left-most object in the top set. Subjects who had rated the similarity of the sets were more likely to choose the right-most object, presumably because both objects were the smallest objects in their sets. Subjects who did not first assess similarity had a tendency to select the middle object, because its size exactly matched the target object’s size. These results are predicted if similarity judgments naturally entail aligning the elements of two scenes. Additional research has shown that relational choices

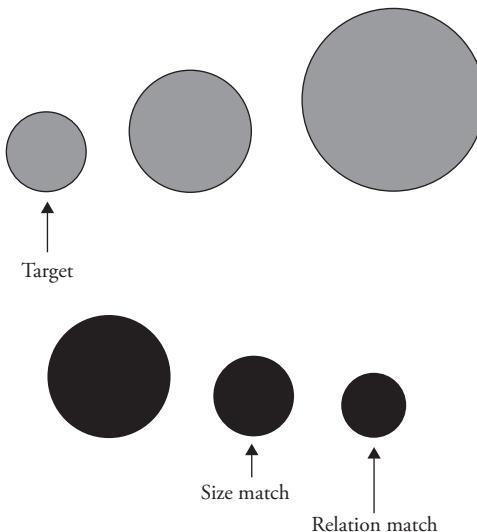


Fig. 10.5 The target from the gray circles could match either the middle black object because they are the same size or the right-most object because both objects are the smallest objects in their sets.

such as “smallest object in its set” tend to influence similarity judgments more than absolute attributes like “3 inches” when the overall amount of relational coherency across sets is high (Goldstone, Medin, & Gentner, 1991), the scenes are superficially sparse rather than rich (Gentner & Rattermann, 1991; Markman & Gentner, 1993a), subjects are given more time to make their judgments (Goldstone & Medin, 1994), the judges are adults rather than children (Gentner & Toupin, 1986), and abstract relations are initially correlated with concrete relations (Kotovsky & Gentner, 1996).

Formal models of alignment-based similarity have been developed to explain how feature matches that belong to well-aligned elements matter more for similarity than matches between poorly aligned elements (Goldstone, 1994a; Larkey & Love, 2003; Love, 2000). Inspired by work in analogical reasoning (Holyoak & Thagard, 1989), Goldstone’s (1994a) SIAM model is a neural network with nodes that represent hypotheses that elements across two scenes correspond to one another. SIAM works by first creating correspondences between the features of scenes. Once features begin to be placed into correspondence, SIAM begins to place objects into correspondence that are consistent with the feature correspondences. Once objects begin to be put into correspondence, activation is fed back down to the feature (mis)matches that are

consistent with the object alignments. In this way, object correspondences influence activation of feature correspondences at the same time that feature correspondences influence the activation of object correspondences. Activation between nodes spreads in SIAM by two principles: (1) nodes that are consistent send excitatory activation to each other and (2) nodes that are inconsistent inhibit each other (see also Holyoak, Chapter 13). Nodes are inconsistent if they create two-to-one alignments—if two elements from one scene would be placed into correspondence with one element of the other scene. Node activations affect similarity via the equation

$$\text{similarity} = \frac{\sum_{i=1}^n (\text{match value}_i * A_i)}{\sum_{i=1}^n A_i},$$

where n is the number of nodes in the system, A_i is the activation of node i , and the match value describes the physical similarity between the two features placed in correspondence according to the node i . By the equation, the influence of a particular matching or mismatching feature across two scenes is modulated by the degree to which the features have been placed in alignment. Consistent with SIAM, (1) aligned feature matches tend to increase similarity more than unaligned feature matches (Goldstone, 1994a), (2) the differential influence between aligned and unaligned feature matches increases as a function of processing time (Goldstone & Medin, 1994), (3) this same differential influences increases with the clarity of the alignments (Goldstone, 1994a), and (4) under some circumstances, adding a poorly aligned feature match can actually decrease similarity by interfering with the development of proper alignments (Goldstone, 1996). Direct tests of SIAM against other models of structural similarity have found advantages for SIAM in that it appropriately weights aligned feature matches more than unaligned feature matches, particularly if the unaligned feature matches are competing against aligned matches (Larkey & Markman, 2005).

Another empirically validated set of predictions stemming from an alignment-based approach to similarity concerns alignable and nonalignable differences (Markman & Gentner, 1993b). Nonalignable differences between two entities are attributes of one entity that have no corresponding attribute in the other entity. Alignable differences are differences

that require that the elements of the entities first be placed in correspondence. When comparing a police car to an ambulance, a nonalignable difference is that police cars have weapons in them, but ambulances do not. There is no clear equivalent of weapons in the ambulance. Alignable differences include the following: Police cars carry criminals to jails rather than carrying sick people to hospitals, a police car is a car while ambulances are vans, and police car drivers are policemen rather than emergency medical technicians. Consistent with the role of structural alignment in similarity comparisons, alignable differences influence similarity more than nonalignable differences do (Markman & Gentner, 1996), and they are more likely to be encoded into memory (Markman & Gentner, 1997). Alignable differences between objects also play a disproportionately large role in distinguishing between different basic-level categories (e.g., cats and dogs) that belong to the same superordinate category (e.g., animals) (Markman & Wisniewski, 1997). In short, knowing these correspondences affects not only how much a matching element increases similarity (Goldstone, 1994a) but also how much a mismatching element decreases similarity.

Thus far, much of the evidence for structural alignment in similarity has used somewhat artificial materials. Often times, the systems describe how “scenes” are compared, with the underlying implication that the elements comprising the scenes are not so tightly connected together as elements comprising objects. Still, if the structural alignment account proves to be fertile, it will be because it is applicable to naturally occurring materials. Toward this goal, researchers have considered structural accounts of similarity in language domains. The confusability of words depends on structural analyses to predict that “stop” is more confusable with “step” than “pest” (the “st” match is in the correct location with “step” but not “pest”), but more confusable with “pest” than “best” (the “p” match counts for something even when it is out of place). Substantial success has been made on the practical problem of determining the structural similarity of words (Bernstein, Demorest, & Eberhardt, 1994; Frisch, Broe, & Pierrehumbert, 1995). Structural alignment has also been implicated when comparing more complex language structures such as sentences (Bassok & Medin, 1997). Likewise, structural similarity has proven to be a useful notion in explaining consumer preferences of commercial products, explaining, for example, why new products are viewed more

favorably when they improve over existing products along alignable rather than unalignable differences (Zhang & Markman, 1998). Additional research has shown that alignment-based models of similarity provide a better account of category-based induction than feature-based models (Lassaline, 1996). Still other researchers have applied structural accounts of similarity to the legal domain (Hahn & Chater, 1998; Simon & Holyoak, 2002; see also Spellman & Schauer, Chapter 36). This area of application is promising because the U.S. legal system is based on cases and precedents, and cases are structurally rich and complex situations involving many interrelated parties. Retrieving a historical precedent, and assessing its relevance to a current case, almost certainly involves aligning representations that are more sophisticated than assumed by geometric or featural models.

Transformational Models

A final historic approach to similarity that has been recently resuscitated is that the comparison process proceeds by transforming one representation into the other. A critical step for these models is to specify what transformational operations are possible.

In an early incarnation of a transformational approach to cognition broadly construed, Garner (1974) stressed the notion of stimuli that are transformationally equivalent and are consequently possible alternatives for each other. In artificial intelligence, Shimon Ullman (1996) has argued that objects are recognized by being aligned with memorized pictorial descriptions. Once an unknown object has been aligned with all candidate models, the best match to the viewed object is selected. The alignment operations rotate, scale, translate, and topographically warp object descriptions. For rigid transformations, full alignment can be obtained by aligning three points on the object with three points on the model description. Unlike recognition strategies that require structural descriptions (e.g., Biederman, 1987; Hummel, 2000, 2001), Ullman’s alignment does not require an image to be decomposed into parts.

In transformational accounts that are explicitly designed to model similarity data, similarity is usually defined in terms of transformational distance. In Wiener-Ehrlich, Bart, and Millward’s (1980) generative representation system, subjects are assumed to possess an elementary set of transformations and to invoke these transformations when analyzing

stimuli. Their subjects saw linear pairs of stimuli such as {ABCD,DABC} or two-dimensional stimuli such as $\left\{ \begin{smallmatrix} AB, DA \\ CD, BC \end{smallmatrix} \right\}$. Subjects were required to rate the similarity of the pairs. The researchers determined transformations that accounted for each subject's ratings from the set {rotate 90 degrees, rotate 180, rotate 270, horizontal reflection, vertical reflection, positive diagonal reflection, negative diagonal reflection}. Similarity was assumed to decrease monotonically as the number of transformations required to make one sequence identical to the other increased.

Imai (1977) made a similar claim. The stimuli used were sequences such as XXOOXXXOXXXXOX where X's represent white ovals, and O's represent black ovals. The four basic transformations were mirror image (XXXXXOO->OOXXXX), phase shift (XXXXXOO->XXXXOOX), reversal (XXXXXOO->OOOOOXX), and wave length (XXOOXXOO->XOXOXOXO). The researcher found that sequences that are two transformations removed (e.g., XXXOXXXOXXXXO and OOXOOOXOOOXO require a phase shift and a reversal to be equated) are rated to be less similar than sequences that can be made identical with one transformation. In addition, sequences that can be made identical by more than one transformation (e.g., XOXOXOXO and OXOXOXOX can be made identical by either mirror image, phase shift, or reversal transformations) are more similar than sequences that have only one identity-producing transformation.

Recent work has followed up on Imai's research and has generalized it to stimulus materials, including arrangements of Lego bricks, geometric complexes, and sets of colored circles (Hahn, Chater, & Richardson, 2003). These researchers have argued that the similarity between two entities is a function of the complexity required to transform the representation of one into the representation of the other. The simpler the transformation, the more similar they are assumed to be. The complexity of a transformation is determined in accord with Kolmogorov complexity theory (Li & Vitanyi, 1997), according to which the complexity of a representation is the length of the shortest computer program that can generate that representation. For example, the conditional Kolmogorov complexity between the sequence 1 2 3 4 5 6 7 8 and 2 3 4 5 6 7 8 9 is small, because the simple instructions that add 1 to each digit (or subtract 1 from each digit) suffice to transform one into the other. Experiments by Hahn

et al. demonstrate that once reasonable vocabularies of transformations are postulated, transformational complexity does indeed predict subjective similarity ratings.

The transformational approach to similarity generally predicts that the similarity of A to B increases as operators become available that can efficiently transform A into B. One interesting empirical prediction from this prediction is that asymmetric similarity judgments may occur if transforming A into B is easier than transforming B into A. Specifically, when participants rated how similar a comparison object was to a reference object, their ratings were higher when the reference object came before, rather than after, the comparison object in a preceding morph sequence animation (Hahn, Close, & Graf, 2009). Transformational models can also account for the pattern of similarity ratings for scenes like that shown in Figure 10.4, if they are given transformational operators such as "delete object," "swap objects," and "move A's texture to B" (Hodgetts, Hahn, & Chater, 2009).

It is useful to compare and contrast alignment-based and transformational accounts of similarity, given that both can account for similarities involving structured scenes like those shown in Figure 10.4. Both approaches place scene elements into correspondence. Whereas the correspondences are explicitly stated in the structural alignment method, they are implicit in transformational alignment. The transformational account often *does* produce globally consistent correspondences, for example correspondences that obey the one-to-one mapping principle, but this consistency is a consequent of applying a pattern-wide transformation and is not enforced by interactions between emerging correspondences. It is revealing that transformational accounts have been applied almost exclusively to perceptual stimuli, whereas structural accounts are often most often applied to conceptual stimuli such as stories, proverbs, and scientific theories (there are also notable structural accounts in perception; e.g., Biederman, 1987; Hummel, 2000; Hummel & Biederman, 1992; Marr & Nishihara, 1978). Defining a set of constrained transformations is much easier for perceptual stimuli. The conceptual similarity between an atom and the solar system could possibly be discovered by transformations. As a start, a miniaturization transformation could be applied to the solar system. However, this single transformation is not nearly sufficient; a nucleus is not simply a small sun. The transformations that would turn the solar system

into an atom are not readily forthcoming. If we allow transformations such as an “earth-becomes-electron” transformation, then we are simply reexpressing the structural alignment approach and its part-by-part alignment of relations and objects.

Some similarity phenomena that are well explained by structural alignment are not easily handled by transformations. To account for the similarity of “BCDCB” and “ABCDCBA,” we could introduce the fairly abstract transformation “Add the left-most letter’s predecessor to both sides of string.” However, the pair “LMN” and “KLMNK” do not seem as similar as the earlier pair, even though the same transformation is applied. A transformation of the form “If the structure is symmetric, then add the preceding element in the series to both ends of the string” presupposes exactly the kind of analysis in defining “symmetric” and “preceding” that are the bread and butter of propositional representations and structural alignment. For this reason, one fertile research direction would be to combine the focus of alignment-based accounts on representing the internal structure within individual scenes with the constraints that transformational accounts provide for establishing psychologically plausible transformations (Hofstadter, 1997; Mitchell, 1993).

Conclusions

A meta-observation about the four models of similarity described earlier is that researchers tend to consider, or in many cases create, entities to be compared that are compatible with their proposed model. Researchers considering dimensional models of similarity tend to choose stimuli that are aptly described in terms of continuous, independent dimensions. Researchers proposing featural models often choose stimuli that have discrete features that either present or absent. Neither researcher tends to choose the richly structured stories or scenes that attract the alignment-based modeler. Researchers proposing transformational accounts often times select stimuli that have easily articulated transformations that can be applied to them. Cases in which a researcher proposing one model is able to accommodate the stimuli used by another modeling tradition are impressive but rare (Hodgetts et al., 2009). In fact, stimulus sets that are well accommodated by, for example, featural models such as additive clustering, are often times poorly accommodated by dimensional approaches such as multidimensional scaling, and vice versa.

Given this situation, ecumenical pluralism is a good strategy. The savvy researcher is well advised

to be proficient with a number of modeling techniques. While the search for a unified model of similarity is tempting and justified by Occam’s razor, it is likely that different kinds of entities are represented differently, and hence enter into similarity comparisons in different ways.

Future Directions

To provide a partial balance to our largely historical focus on similarity, we will conclude by raising some unanswered questions for the field. These questions are rooted in a desire to connect the study of similarity to cognition as a whole.

Is Similarity Flexible Enough to Provide Useful Explanations of Cognition?

The study of similarity is typically justified by the argument that so many theories in cognition depend upon similarity as a theoretical construct. An account of what makes problems, memories, objects, and words similar to one another often provides the backbone for our theories of problem solving, attention, perception, and cognition. As William James put it, “This sense of Sameness is the very keel and backbone of our thinking” (James, 1890/1950, p. 459).

However, others have argued that similarity is not flexible enough to provide a sufficient account, although it may be a necessary component. There have been many empirical demonstrations of apparent dissociations between similarity and other cognitive processes, most notably categorization. Researchers have argued that cognition is frequently based on theories (Murphy & Medin, 1985), rules (Sloman, 1996; Smith & Sloman, 1994), or strategies that go beyond “mere” similarity. To take an example from Murphy and Medin (1985), consider a man jumping into a swimming pool fully clothed. This man may be categorized as drunk because we have a theory of behavior and inebriation that explains the man’s action. Murphy and Medin argue that the categorization of the man’s behavior does not depend on matching the man’s features to the category *drunk*’s features. It is highly unlikely that the category *drunk* would have such a specific feature as “jumps into pools fully clothed.” It is not the similarity between the instance and the category that determines the instance’s classification; it is the fact that our category provides a theory that explains the behavior.

Developmental psychologists have argued that even young children have inchoate theories that allow them to go beyond superficial similarities

in creating categories (Carey, 1985; Gelman & Markman, 1986; Keil, 1989; see Gelman & Frazier, Chapter 26). For example, Carey (1985) observed that children choose a toy monkey over a worm as being more similar to a human; but that when they are told that humans have spleens, children are more likely to infer that the worm has a spleen than that the toy monkey does. Thus, the categorization of objects into “spleen” and “no spleen” groups does not appear to depend on the same knowledge that guides similarity judgments. Adults show similar dissociations between similarity and categorization. In an experiment by Rips (1989), an animal that is transformed (by toxic waste) from a bird into something that looks like an insect is judged by subjects to be more similar to an insect, but it is still judged to *be* a bird. Again, the category judgment seems to depend on biological, genetic, and historical knowledge, while the similarity judgments seem to depend more on gross visual appearance (see also Keil, 1989; Rips & Collins, 1993; Rips et al., Chapter 11).

Despite the growing body of evidence that similarity appraisals do not always track categorization decisions, there are still some reasons to be sanguine about the continued explanatory relevance of similarity. Categorization itself may not be completely flexible. People are influenced by similarity despite the subjects’ intentions and the experimenters’ instructions (Smith & Sloman, 1994). Allen and Brooks (1991) gave subjects an easy rule for categorizing cartoon animals into two groups. Subjects were then transferred to the animals that looked very similar to one of the training stimuli but belonged in a different category. These animals were categorized more slowly and less accurately than animals that were equally similar to an old animal but also belonged in the same category as the old animal. Likewise, Palmeri (1997) showed that even for the simple task of counting the number of dots, subjects’ performance is improved when a pattern of dots is similar to a previously seen pattern with the same numerosity and worse when the pattern is similar to a previously seen pattern with different numerosity. People seem to have difficulties ignoring similarities between old and new patterns, even when they know a straightforward and perfectly accurate categorization rule.

There may be a mandatory consideration of similarity in many categorization judgments (Goldstone, 1994b), adding constraints to categorization. At the same time, similarity may be more flexible and sophisticated than commonly acknowledged (Jones & Smith, 1993), bridging the gap between

similarity and high-level cognition. Krumhansl (1978) argued that similarity between objects decreases when they are surrounded by many close neighbors that were also presented on previous trials (also see Wedell, 1994). Tversky (1977) obtained evidence for an *extension effect*, according to which features influence similarity judgments more when they vary within an entire set of stimuli. Items presented within a particular trial also influence similarity judgments. Perhaps the most famous example of this is Tversky’s (1977) *Diagnosticity effect*, according to which features that are diagnostic for relevant classifications will have disproportionate influence on similarity judgments. More recently, Medin, Goldstone, and Gentner (1993) have argued that different comparison standards are created depending on the items that are present on a particular trial. Other research has documented intransitivities in similarity judgments, situations where A is judged to be more similar to T than is B, B is more similar to T than is C, and C is more similar to T than is A (Goldstone, Medin, & Halberstadt, 1997). This type of finding also suggests that the properties used to assess the similarity of objects are determined, in part, by the compared objects themselves.

Similarity judgments not only depend on the context established by recently exposed items, simultaneously presented items, and inferred contrast sets but also on the observer. Suzuki, Ohnishi, and Shigemasu (1992) have shown that similarity judgments depend on level of expertise and goals. Expert and novice subjects were asked to solve the Tower of Hanoi puzzle and judge the similarity between the goal and various states. Experts’ similarity ratings were based on the number of moves required to transform one position to the other. Less expert subjects tended to base their judgments on the number of shared superficial features. Similarly, Hardiman, Dufresne, and Mestre (1989) found that expert and novice physicists evaluate the similarity of physics problems differently, with experts basing similarity judgments more on general principles of physics than on superficial features (see Sjoberg, 1972 for other expert/novice differences in similarity ratings). The dependency of similarity on observer-, task-, and stimulus-defined contexts offers the promise that it is indeed flexible enough to subserve cognition.

Is Similarity Too Flexible to Provide Useful Explanations of Cognition?

As a response to the skeptic of similarity’s usefulness, the preceding two paragraphs could have the

exact opposite of their intended effect. The skeptic might now feel that similarity is much *too* flexible to be a stable ground for cognition. In fact, Nelson Goodman (1972) has put forth exactly this claim, maintaining that the notion of similarity is either vague or unnecessary. He argued that “when to the statement that two things are similar we add a specification of the property that they have in common...we render it [the similarity statement] superfluous” (p. 445). That is, all of the potential explanatory work is done by the “with respect to property Z” clause and not by the similarity statement. Instead of saying, “This object belongs to Category A because it is similar to A items with respect to the property ‘red,’ “ we can simplify matters by removing any notion of similarity with “This object belongs to Category A because it is red.”

There are reasons to resist Goodman’s conclusion that “Similarity tends under analysis either to vanish entirely or to require for its explanation just what it purports to explain” (p. 446). In most cases, similarity is useful precisely because we cannot flesh out the “respect to property Z” clause with just a single property. Evidence suggests that assessments of overall similarity are natural and perhaps even “primitive.” Evidence from children’s perception of similarity suggests that children are particularly likely to judge similarity on the basis of many integrated properties rather than analysis into dimensions. Even dimensions that are perceptually separable are treated as fused in similarity judgments (Smith & Kemler, 1978). Children under 5 years of age tend to classify on the basis of overall similarity and not on the basis of a single criterial attribute (Keil, 1989; Smith, 1989). Children often have great difficulty identifying the dimension along which two objects vary, even though they can easily identify that the objects are different in some way (Kemler, 1983). Smith (1989) argued that it is relatively difficult for young children to say whether two objects are identical on a particular property, but relatively easy for them to say whether they are similar across many dimensions.

There is also evidence that adults often have an overall impression of similarity without analysis into specific properties. Ward (1983) found that adult subjects who tended to group objects quickly also tended to group objects like children, by considering overall similarity across all dimensions instead of maximal similarity on one dimension. Likewise, Smith and Kemler (1984) found that adults who were given a distracting task produced more

judgments by overall similarity than subjects who were not. To the extent that similarity is determined by many properties, it is less subject to drastic context-driven changes. Furthermore, integrating multiple sources of information into a single assessment of similarity becomes particularly important. The four approaches to similarity described in the previous section all provide methods for integrating multiple properties into a single similarity judgment, and as such, they go significantly beyond simply determining a single “property Z” to attend to.

A final point to make about the potential overflexibility of similarity is that although impressions of similarity can change with context and experience, automatic and “generic” assessments of similarity typically change slowly and with considerable inertia. Similarities that were once effortful and strategic become second nature to the organism. Roughly speaking, this is the process of *perceiving* what was once a *conceptual* similarity. At first, the novice mycologist explicitly uses rules for perceiving the dissimilarity between the pleasing Agaricus Bisporus mushroom and the deadly Amanita Phalloides. With time, this dissimilarity ceases to be effortful and rule based, and becomes perceptual and phenomenologically direct. When this occurs, the similarity becomes generic and default, and it can be used as the ground for new strategic similarities. In this way, our cognitive abilities gradually attain sophistication, treating territory as level ground that once made for difficult mental climbing. A corollary of this contention is that our default impression of similarity does not typically mislead us; it is explicitly *designed* to lead us to see relations between things that often function similarly in our world. People, with good reason, expect their default similarity assessments to provide good clues about where to uncover directed, nonapparent similarities (Medin & Ortony, 1989).

Should “Similarity” Even Be a Field of Study Within Cognitive Science?

This survey has proceeded under the convenient fiction that it is possible to tell a general story for how people compare things. One reason to doubt this assumption is that the methods used for assessing similarity have large effects on the resulting similarity viewed. Similarity as measured by ratings is not equivalent to similarity as measured by perceptual discriminability. Although these measures correlate highly, systematic differences are found (Podgorny & Garner, 1979; Sergent & Takane, 1987). For

example, Beck (1966) found that an upright T is rated as more similar to a tilted T than an upright L, but that it is also more likely to be perceptually grouped with the upright L's. Previously reviewed experiments indicate the nonequivalence of assessments that use similarity versus dissimilarity ratings, categorization versus forced-choice similarity judgments, or speeded versus leisurely judgments. In everyday discourse we talk about the similarity of two things, forgetting that this assessment depends upon a particular task and circumstance.

Furthermore, it may turn out that the calculation of similarity is fundamentally different for different domains (see Medin, Lynch, & Solomon, 2000, for a thoughtful discussion of this issue). To know how to calculate the similarity of two faces, one would need to study faces specifically, and the eventual account need not inform researchers interested in the similarity of words, works of music, or trees. A possible conclusion is that similarity is not a coherent notion at all. The term *similarity*, like the terms *bug* or *family values*, may not pick out a consolidated or principled set of things.

Although we sympathize with the impulse toward domain-specific accounts of similarity, we also believe in the value of studying general principles of comparison that potentially underlie many domains. Although we do not know whether general principles exist, one justification for pursuing them is the large payoff that would result from discovering these principles if they *do* exist. A historically fruitful strategy, exemplified by Einstein's search for a law to unify gravitational and electromagnetic acceleration, and Darwin's search for a unified law to understand the origins of humans and other animals, has been to understand differences as parametric variations within a single model. Finding differences across tasks does not necessarily indicate the incoherency of similarity. An alternative perspective would use these task differences as an illuminating source of information in developing a unified account. The systematic nature of these task differences should stimulate accounts that include a formal description not only of stimulus components but also of task components. Future success in understanding the task of comparison may depend on comparing tasks.

Acknowledgments

This research was funded by National Science Foundation REESE grant 0910218 and Department of Education IES grant R305A1100060.

References

- Aha, D. W. (1992). Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. *International Journal of Man Machine Studies*, 36, 267–287.
- Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General*, 120, 3–19.
- Attneave, F. (1950). Dimensions of similarity. *American Journal of Psychology*, 63, 516–556.
- Bassok, M., & Medin, D. L. (1997). Birds of a feather flock together: Similarity judgments with semantically rich stimuli. *Journal of Memory & Language*, 36, 311–336.
- Beck, J. (1966). Effect of orientation and of shape similarity on perceptual grouping. *Perception and Psychophysics*, 1, 300–302.
- Bernstein, L. E., Demorest, M. E., & Eberhardt, S. P. (1994). A computational approach to analyzing sentential speech perception: Phoneme-to-phoneme stimulus/response alignment. *Journal of the Acoustical Society of America*, 95, 3617–3622.
- Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94, 115–147.
- Burgess, C., & Lund, K. (2000). The dynamics of meaning in memory. In E. Dietrich & A. B. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines*. (pp. 117–156). Mahwah, NJ: Erlbaum.
- Bush, R. R., & Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological Review*, 58, 413–423.
- Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: Bradford Books.
- Carroll, J. D., & Wish, M. (1974). Models and methods for three-way multidimensional scaling. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 2, pp. 57–105). San Francisco, CA: Freeman.
- Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121–152.
- Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8, 240–247.
- Corter, J. E. (1987). Similarity, confusability, and the density hypothesis. *Journal of Experimental Psychology: General*, 116, 238–249.
- Corter, J. E. (1988). Testing the density hypothesis: Reply to Krumhansl. *Journal of Experimental Psychology: General*, 117, 105–106.
- Edelman, S. (1999). *Representation and recognition in vision*. Cambridge, MA: MIT Press.
- Eisler, H., & Ekman, G. (1959). A mechanism of subjective similarity. *Acta Psychologica*, 16, 1–10.
- Estes, W. K. (1994). *Classification and cognition*. New York: Oxford University Press.
- Falkenhainer, B., Forbus, K.D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41, 1–63.
- Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press/Bradford Books.
- Frisch, S. A., Broe, M. B., & Pierrehumbert, J. B. (1995). The role of similarity in phonology: Explaining OCP-Place. In K. Elenius & P. Branderud (Eds.), *Proceedings of the 13th International Conference of the Phonetic Sciences*, 3, 544–547.

- Gardenfors, P. (2000). *Conceptual spaces: The geometry of thought*. Cambridge, MA: MIT Press.
- Garner, W. R. (1974). *The processing of information and structure*. New York: Wiley.
- Gati, I., & Tversky, A. (1984). Weighting common and distinctive features in perceptual and conceptual judgments. *Cognitive Psychology*, 16, 341–370.
- Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, 23, 183–209.
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155–170.
- Gentner, D., & Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman & J. P. Byrnes (Eds.), *Perspectives on language and thought interrelations in development* (pp. 225–277). Cambridge, England: Cambridge University Press.
- Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science*, 10(3), 277–300.
- Gilmore, G. C., Hersh, H., Caramazza, A., & Griffin, J. (1979). Multidimensional letter similarity derived from recognition errors. *Perception and Psychophysics*, 25, 425–431.
- Gluck, M. A. (1991). Stimulus generalization and representation in adaptive network models of category learning. *Psychological Science*, 2, 50–55.
- Gluck, M. A., & Bower, G. H. (1990). Component and pattern information in adaptive networks. *Journal of Experimental Psychology: General*, 119, 105–109.
- Goldstone, R. L. (1994a). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 3–28.
- Goldstone, R. L. (1994b). The role of similarity in categorization: Providing a groundwork. *Cognition*, 52, 125–157.
- Goldstone, R. L. (1996). Alignment-based nonmonotonicities in similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 988–1001.
- Goldstone, R. L., & Medin, D. L. (1994). The time course of comparison. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 29–50.
- Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relations, attributes, and the non-independence of features in similarity judgments. *Cognitive Psychology*, 222–264.
- Goldstone, R. L., Medin, D. L., & Halberstadt, J. (1997). Similarity in context. *Memory and Cognition*, 25, 237–255.
- Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Problems and projects* (pp. 437–446). New York: The Bobbs-Merrill Co.
- Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. T. (2007). Topics in semantic representation. *Psychological Review*, 114(2), 211–244.
- Hahn, U. (2003). Similarity. In L. Nadel (Ed.), *Encyclopedia of cognitive science* (pp. 386–388). London: Macmillan.
- Hahn, U., & Chater, N. (1998). Understanding similarity: A joint project for psychology, case-based reasoning and law. *Artificial Intelligence Review*, 12, 393–427.
- Hahn, U., Chater, N., & Richardson, L. B. (2003). Similarity as transformation. *Cognition*, 87, 1–32.
- Hahn, U., Close, J., & Graf, M. (2009). Transformation direction influences shape similarity judgments. *Psychological Science*, 20, 447–454.
- Hardiman, P. T., Dufresne, R., & Mestre, J. P. (1989). The relation between problem categorization and problem solving among experts and novices. *Memory and Cognition*, 17, 627–638.
- Hayes-Roth, B., & Hayes-Roth, F. (1977). Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning and Verbal Behavior*, 16, 321–338.
- Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 411–422.
- Hodgetts, C. J., Hahn, U., & Chater, N. (2009). Transformation and alignment in similarity. *Cognition*, 113, 62–79.
- Hofstadter, D. (1997). *Fluid concepts and creative analogies: computer models of the fundamental mechanisms of thought*. New York: Basic Books.
- Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: Bradford Books/MIT Press.
- Holyoak, K. J., & Gordon, P. C. (1983). Social reference points. *Journal of Personality and Social Psychology*, 44, 881–887.
- Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory and Cognition*, 15, 332–340.
- Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science*, 13, 295–355.
- Holyoak, K. J., & Hummel, J. E. (2000). The proper treatment of symbols in a connectionist architecture. In E. Dietrich & A. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines* (pp. 229–263). Hillsdale, NJ: Erlbaum.
- Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science*, 13, 295–355.
- Holyoak, K. J., & Thagard, P. (1995). Mental leaps: Analogy in creative thought. Cambridge, MA: MIT Press.
- Horgan, D. D., Millis, K., & Neimeyer, R. A. (1989). Cognitive reorganization and the development of chess expertise. *International Journal of Personal Construct Psychology*, 2, 15–36.
- Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, 195, 215–243.
- Hummel, J. E. (2000). Where view-based theories break down: The role of structure in shape perception and object recognition. In E. Dietrich & A. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines* (pp. 157–185). Hillsdale, NJ: Erlbaum.
- Hummel, J. E. (2001). Complementary solutions to the binding problem in vision: Implications for shape perception and object recognition. *Visual Cognition*, 8, 489–517.
- Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, 99, 480–517.
- Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104, 427–466.
- Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, 110, 220–263.
- Imai, S. (1977). Pattern similarity and cognitive transformations. *Acta Psychologica*, 41, 433–447.
- James, W. (1890/1950). *The principles of psychology*. New York: Dover. (Original work published 1890).
- Jakobson, R., Fant, G., & Halle, M. (1963). *Preliminaries to speech analysis: The distinctive features and their correlates*. Cambridge, MA: MIT Press.

- Jones, S. S., & Smith, L. B. (1993). The place of perception in children's concepts. *Cognitive Development*, 8, 113–139.
- Katz, J. J., & Fodor, J. (1963). The structure of semantic theory. *Language*, 39, 170–210.
- Keil, F. C. (1989). *Concepts, kinds and development*. Cambridge, MA: Bradford Books/MIT Press.
- Kemler, D. G. (1983). Holistic and analytic modes in perceptual and cognitive development. In T. J. Tighe & B. E. Shepp (Eds.), *Perception, cognition, and development: Interactional analyses*. (pp. 77–101). Hillsdale, NJ: Erlbaum.
- Kohonen, T. (1995). Self-organizing maps. Berlin: Springer-Verlag.
- Kolers, P. A., & Roediger, H. L. (1984). Procedures of mind. *Journal of Verbal Learning and Verbal Behavior*, 23, 425–449.
- Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, 67, 2797–2822.
- Krumhansl, C. L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review*, 85, 450–463.
- Krumhansl, C. L. (1988). Testing the density hypothesis: Comment on Carter. *Journal of Experimental Psychology: General*, 117, 101–104.
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99, 22–44.
- Lamberts, K. (2000). Information-accumulation theory of speeded categorization. *Psychological Review*, 107, 227–260.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211–240.
- Larkey, L. B., & Love, B. C. (2003). CAB: Connectionist analogy builder. *Cognitive Science*, 27, 781–794.
- Larkey, L. B., & Markman, A. B. (2005). Processes of similarity judgment. *Cognitive Science*, 29, 1061–1076.
- Lassaline, M. E. (1996). Structural alignment in induction and similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 754–770.
- Lee, M. D. (1998). Neural feature abstraction from judgments of similarity. *Neural Computation*, 10(7), 1815–1830.
- Lee, M. D. (2001). Determining the dimensionality of multidimensional scaling representations for cognitive modeling. *Journal of Mathematical Psychology*, 45, 149–166.
- Lee, M. D. (2002a). A simple method for generating additive clustering models with limited complexity. *Machine Learning*, 49, 39–58.
- Lee, M. D. (2002b). Generating additive clustering models with limited stochastic complexity. *Journal of Classification*, 19, 69–85.
- Li, M., & Vitanyi, P. (1997). *An introduction to Kolmogorov complexity and its applications*, (2nd ed.). New York: Springer-Verlag.
- Love, B. C. (2000). A computational level theory of similarity. In *Proceeding of the Cognitive Science Society* (pp. 316–321). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Markman, A. B., & Gentner, D. (1993a). Structural alignment during similarity comparisons. *Cognitive Psychology*, 25, 431–467.
- Markman, A. B., & Gentner, D. (1993b). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language*, 32, 517–535.
- Markman, A. B., & Gentner, D. (1996). Commonalities and differences in similarity comparisons. *Memory and Cognition*, 24, 235–249.
- Markman, A. B., & Gentner, D. (1997). The effects of alignability on memory. *Psychological Science*, 8, 363–367.
- Markman, A. B., & Wisniewski, E. J. (1997). Similar and different: The differentiation of basic-level categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 54–70.
- Marr, D. (1982). *Vision*. San Francisco, CA: Freeman.
- Marr, D., & Nishihara, H. K. (1978). Representation and recognition of three dimensional shapes. *Proceedings of the Royal Society of London: Series B*, 200, 269–294.
- Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, 100, 254–278.
- Medin, D. L., LynChapter E. B., & Solomon, K. O. (2000). Are there kinds of concepts? *Annual Review of Psychology*, 51, 121–147.
- Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–195). Cambridge, MA: Cambridge University Press.
- Medin, D. L., & Schaffer, M. M. (1978). A context theory of classificationlearning. *Psychological Review*, 85, 207–238.
- Mitchell, M. (1993). Analogy-making as perception: a computer model. Cambridge, MA: MIT Press.
- Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.
- Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92, 289–316.
- Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.
- Navarro, D. J., & Griffiths, T. L. (2008). Latent features in similarity judgments: A nonparametric Bayesian approach. *Neural Computation*, 20, 2597–2628.
- Navarro, D. J., & Lee, M. D. (2003). Combining dimensions and features in similarity based representations. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 67–74). Cambridge, MA: MIT Press.
- Navarro, D. J., & Perfors, A. F. (2010). Similarity, feature discovery and the size principle. *Acta Psychologica*, 133, 256–268.
- Nickerson, R. S. (1972). Binary classification reaction time: A review of some studies of human information-processing capabilities. *Psychonomic Monograph Supplements*, 4(6), 275–317.
- Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10, 104–114.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115, 39–57.
- Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology*, 23, 94–140.
- Osterholm, K., Woods, D. J., & Le Unes, A. (1985). Multidimensional scaling of Rorschach inkblots: Relationships with structured self-report. *Personality and Individual Differences*, 6, 77–82.
- Palmer, S. E. (1975). Visual perception and world knowledge. In D. A. Norman & D. E. Rumelhart (Eds.), *Explorations in cognition* (pp. 214–246). San Francisco, CA: Freeman.
- Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 324–354.

- Podgorny P., & Garner, W. R. (1979). Reaction time as a measure of inter-intraobject visual similarity: Letters of the alphabet. *Perception and Psychophysics*, 26, 37–52.
- Polk, T. A., Behensky, C., Gonzalez, R., & Smith, E. E. (2002). Rating the similarity of simple perceptual stimuli: Asymmetries induced by manipulating exposure frequency. *Cognition*, 82, B75–B88.
- Plyshyn, Z. W. (1985). *Computation and cognition*. Cambridge, MA: MIT press.
- Quine, W. V. (1969). *Ontological relativity and other essays*. New York: Columbia University Press.
- Quine, W. V. (1977). Natural kinds. In S. P. Schwartz (Ed.), *Naming, necessity, and natural kinds* (pp. 155–177). Ithaca, NY: Cornell University Press.
- Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, 88, 93–134.
- Richardson, M. W. (1938). Multidimensional psychophysics. *Psychological Bulletin*, 35, 659–660.
- Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge, England: Cambridge University Press.
- Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity, analogy, and thought* (pp. 21–59). Cambridge, England: Cambridge University Press.
- Rips, L. J., & Collins, A. (1993). Categories and resemblance. *Journal of Experimental Psychology: General*, 122, 468–486.
- Ritov, I., Gati, I., & Tversky, A. (1990). Differential weighting of common and distinctive components. *Journal of Experimental Psychology: General*, 119, 30.
- Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 629–639.
- Ross, B. H. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 456–468.
- Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, 26, 521–562.
- Schank, R. C. (1982). *Dynamic memory: A theory of reminding and learning in computers and people*. Cambridge, England: Cambridge University Press.
- Schvaneveldt, R. (1985). Measuring the structure of expertise. *International Journal of Man-Machine Studies*, 23, 699–728.
- Sergent, J., & Takane, Y. (1987). Structures in two-choice reaction-time data. *Journal of Experimental Psychology: Human Perception and Performance*, 13, 300–315.
- Shepard, R. N. (1962a). The analysis of proximities: Multidimensional scaling with an unknown distance function. Part I. *Psychometrika*, 27, 125–140.
- Shepard, R. N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. Part II. *Psychometrika*, 27, 219–246.
- Shepard, R. N. (1972). Psychological representation of speech sounds. In E. E. David, Jr. & P. B. Denes (Eds.), *Human communication: A unified view* (pp. 165–173). New York: McGraw-Hill.
- Shepard, R. N. (1982). Geometrical approximations to the structure of musical pitch. Chapter *Psychological Review*, 89, 305–333.
- Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237, 1317–1323.
- Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, 86, 87–123.
- Simon, D., & Holyoak, K. J. (2002). Structural dynamics of cognition: From consistency theories to constraint satisfaction. *Personality and Social Psychology Review*, 6, 283–294.
- Sjöberg, L. (1972). A cognitive theory of similarity. *Goteborg Psychological Reports*, 2(10).
- Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, 25, 231–280.
- Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119, 3–22.
- Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, 81, 214–241.
- Smith, E. E., & Sloman, S. A. (1994). Similarity-versus rule-based categorization. *Memory and Cognition*, 22, 377–386.
- Smith, J. D., & Kemler, D. G. (1984). Overall similarity in adults' classification: The child in all of us. *Journal of Experimental Psychology: General*, 113, 137–159.
- Smith, L. B. (1989). From global similarity to kinds of similarity: The construction of dimensions in development. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 146–178). Cambridge, England: Cambridge University Press.
- Smith, L. B., & Kemler, D. G. (1978). Levels of experienced dimensionality in children and adults. *Cognitive Psychology*, 10, 502–532.
- Suzuki, H., Ohnishi, H., & Shigemasu, K. (1992). Goal-directed processes in similarity judgment. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 343–348). Hillsdale, NJ: Erlbaum.
- Tarr, M. J., & Gauthier, I. (1998). Do viewpoint-dependent mechanisms generalize across members of a class? [Special issue: Image-based object recognition in man, monkey, and machine]. *Cognition*, 67, 73–110.
- Tenenbaum, J. B. (1996). Learning the structure of similarity. In G. Tesauro, D. S. Touretzky, & T. K. Leen (Eds.), *Advances in neural information processing systems 8* (pp. 4–9). Cambridge, MA: MIT Press.
- Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in neural information processing systems 11* (pp. 59–65). Cambridge, MA: MIT Press.
- Tenenbaum, J. B., De Silva, V., & Lanford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290, 22–23.
- Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity and Bayesian inference. *Behavioral and Brain Sciences*, 24, 629–640.
- Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.
- Torgerson, W. S. (1965). Multidimensional scaling of similarity. *Psychometrika*, 30, 379–393.
- Treisman, A. M. (1986). Features and objects in visual processing. *Scientific American*, 255, 106–115.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327–352.
- Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review*, 89, 123–154.
- Tversky, A., & Hutchinson, J. W. (1986). Nearest neighbor analysis of psychological spaces. *Psychological Review*, 93, 3–22.

- Ullman, S. (1996). *High-level vision: Object recognition and visual cognition*. London: MIT Press.
- Ward, T. B. (1983). Response tempo and separable-integral responding: Evidence for an integral-to-separable processing sequence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance*, 9, 103–112.
- Wedell, D. (1994). Context effects on similarity judgments of multidimensional stimuli: Inferring the structure of the emotion space. *Journal of Experimental Social Psychology*, 30, 1–38.
- Wiener-Ehrlich, W. K., Bart, W. M., & Millward, R. (1980). An analysis of generative representation systems. *Journal of Mathematical Psychology*, 21(3), 219–246.
- Zhang, S., & Markman, A. B. (1998). Overcoming the early entrant advantage: The role of alignable and non-alignable differences. *Journal of Marketing Research*, 35, 413–426.

Concepts and Categories: Memory, Meaning, and Metaphysics

Lance J. Rips, Edward E. Smith, and Douglas L. Medin

Abstract

The psychological study of concepts has two main goals: explaining how people's knowledge of categories such as tables or cats enables them to classify or recognize members of those categories, and explaining how knowledge of word meanings (e.g., the meaning of *table* and *cat*) enables people to make inferences and to compute the meanings of phrases and sentences. We review current theories and data relevant to these two functions of concepts, including recent insights from cognitive neuropsychology. Both kinds of theories have evolved in ways that suggest that people make use of mental representations at several levels of complexity, from sparse, atomic concepts to complex, knowledge-intensive ones. We examine the implications of this variety for issues including psychological essentialism and domain specificity.

Key Words: concepts, categories, semantics, prototypes, exemplars, generics, polysemy, essentialism, sortals, folk biology

Introduction

The concept of concepts is difficult to define, but no one doubts that concepts are fundamental to mental life and human communication. Cognitive scientists generally agree that a concept is a mental representation. Theories in psychology have concentrated on concepts that pick out certain sets of entities: categories. That is, concepts *refer*, and what they refer to are categories. It is also commonly assumed that category membership is not arbitrary but rather a principled matter. What goes into a category belongs there by virtue of some law-like regularities. But beyond these sparse facts, the concept CONCEPT is up for grabs. As an example, suppose you have the concept TRIANGLE represented as "a closed geometric form having three sides." In this case, the concept is a definition. But it is unclear what else might be in your triangle concept. Does it include the fact that geometry books discuss them (though some don't) or that they have 180 degrees (though in hyperbolic geometry none do)? It is also

unclear how many concepts have definitions or what substitutes for definitions in ones that don't.

Our goal in this chapter is to provide an overview of work on concepts and categories in the last half century. There has been such a consistent stream of research over this period that one critic of this literature, Gregory Murphy (2002), felt compelled to call his monograph, *The Big Book of Concepts*. Our task is eased by recent reviews, including Murphy's aptly named one (e.g., Medin, Lynch & Solomon, 2000; Murphy, 2002; Rips, 2001; Wisniewski, 2002). Their thoroughness gives us the luxury of doing a review focused on a single perspective—the relations among concepts, memory, and meaning.

The remainder of this chapter is organized as follows. In the rest of this section, we briefly describe some of the tasks or functions that cognitive scientists have expected concepts to perform. This will provide a roadmap to important lines of research on concepts and categories. Next, we return to developments in the late 1960s and early 1970s that raised

the exciting possibility that laboratory studies could provide deep insights into both concept representations and the organization of (semantic) memory. Then we describe the sudden collapse of this optimism and the ensuing lines of research that, however intriguing and important, ignored questions about semantic memory. Next we trace a number of relatively recent developments under the somewhat whimsical heading “Psychometaphysics.” This is the view that concepts are embedded in (perhaps domain-specific) theories. This will set the stage for returning to the question of whether research on concepts and categories is relevant to semantics and memory organization. We will use that question to speculate about future developments in the field. In this review, we will follow the usual conventions of using words in all caps to refer to concepts and quoted words to refer to linguistic expressions.

Functions of Concepts

For purposes of this review, we will collapse the many ways people can use concepts into two broad functions: categorization and communication. The conceptual function that most research has targeted is *categorization*, the process by which mental representations (concepts) determine whether some entity is a member of a category. Categorization enables a wide variety of subordinate functions because classifying something as a category member allows people to bring their knowledge of the category to bear on the new instance. Once people categorize some novel entity, they can use relevant knowledge for *understanding* and *prediction*. Recognizing a cylindrical object as a flashlight allows you to understand its parts, trace its functions, and predict its behavior. For example, you can confidently infer that the flashlight will have one or more batteries, will have some sort of switch, and will normally produce a beam of light when the switch is pressed.

Not only do people categorize in order to understand new entities, they also use the new entities to modify and update their concepts. In other words, categorization supports *learning*. Encountering a member of a category with a novel property—for example, a flashlight that has a siren for emergencies—can result in that novel property being incorporated into the conceptual representation. Relations between categories may also support inference and learning. For example, finding out that flashlights can contain sirens may lead you to entertain the idea that cell phones and fire extinguishers might also contain sirens. Hierarchical

conceptual relations support both inductive and deductive *reasoning*. If you know all trees contain xylem and hawthorns are trees, then you can deduce that hawthorns contain xylem. In addition, finding out that white oaks contain phloem provides some support for the inductive inference that other kinds of oaks contain phloem. People also use categories to instantiate goals in *planning* (Barsalou, 1983). For example, a person planning to do some night fishing might create an ad hoc concept, THINGS TO BRING ON A NIGHT FISHING TRIP, which would include a fishing rod, tackle box, mosquito repellent, and a flashlight.

Concepts are also centrally involved in *communication*. Many of our concepts correspond to lexical entries, such as the English word “flashlight.” In order for people to avoid misunderstandings, they must have comparable concepts in mind. If A’s concept of cell phone corresponds with B’s concept of flashlight, it won’t go well if A asks B to make a call. An important part of the function of concepts in communication is their ability to combine in order to create an unlimited number of new concepts. Nearly every sentence you encounter is new—one you’ve never heard or read before—and concepts (along with the sentence’s grammar) must support your ability to understand it. Concepts are also responsible for more creative uses of language. For example, from the base concepts of TROUT and FLASHLIGHT, you might create a new concept, TROUT FLASHLIGHT, which in the context of our current discussion would presumably be a flashlight used when trying to catch trout (and not a flashlight with a picture of a trout on it, though this may be the correct interpretation in some other context). A major research challenge is to understand the principles of *conceptual combination* and how they relate to communicative contexts (see Fodor, 1994, 1998; Gleitman & Papafragou, Chapter 28; Hampton, 1997; Partee, 1995; Rips, 1995; Wisniewski, 1997).

Overview

So far, we have introduced two roles for concepts: categorization (broadly construed) and communication. These functions and associated subfunctions are important to bear in mind because studying any one in isolation can lead to misleading conclusions about conceptual structure (see Solomon, Medin, & Lynch, 1999, for a review bearing on this point). At this juncture, however, we need to introduce one more plot element into the story we are telling. Presumably everything we have been talking about has implications for

human memory and memory organization. After all, concepts are mental representations, and people must store these representations somewhere in memory. However, the relation between concepts and memory may be more intimate. A key part of our story is what we call “the semantic memory marriage,” the idea that memory organization corresponds to meaningful relations between concepts. Mental pathways that lead from one concept to another—for example, from ELBOW to ARM—represent relations like IS A PART OF that link the same concepts. Moreover, these memory relations may supply the concepts with all or part of their meaning. By studying how people use concepts in categorizing and reasoning, researchers could simultaneously explore memory structure and the structure of the mental lexicon. In other words, the idea behind this move was to unify categorization, communication (in its semantic aspects), and memory organization. As we’ll see, this marriage was somewhat troubled, and there are many rumors about its breakup. But we are getting ahead of our story. The next section begins with the initial romance.

A Mini-History

Research on concepts in the middle of the last century reflected a gradual easing away from behaviorist and associative learning traditions. The focus, however, remained on learning. Most of this research was conducted in laboratories using artificial categories (a sample category might be any geometric figure that is both red and striped) and directed at one of two questions: (a) Are concepts learned by gradual increases in associative strength, or is learning all or none (Levine, 1971; Trabasso & Bower, 1968)? and (b) Which kinds of rules or concepts (e.g., disjunctive, such as RED OR STRIPED, versus conjunctive, such as RED AND STRIPED) are easiest to learn (Bruner, Goodnow, & Austin, 1956; Bourne, 1970; Restle, 1962)?

This early work tended either to ignore real-world concepts (Bruner et al., 1956, represent something of an exception here) or to assume implicitly that real-world concepts are structured according to the same kinds of rules that defined the artificial ones. According to this tradition, category learning is equivalent to identifying the definitions that determine category membership.

Early Theories of Semantic Memory

Although the work on rule learning set the stage for what was to follow, two developments associated

with the emergence of cognitive psychology dramatically changed how people thought about concepts.

TURNING POINT I: MODELS OF MEMORY ORGANIZATION

The idea of programming computers to do intelligent things (artificial intelligence or AI) had an important influence on the development of new approaches to concepts. Quillian (1967) proposed a hierarchical model for storing semantic information in a computer that was quickly evaluated as a candidate model for the structure of human memory (Collins & Quillian, 1969). Figure 11.1 provides an illustration of part of a memory hierarchy that is similar to what the Quillian model suggests.

First, note that the network follows a principle of cognitive economy. Properties true of all animals, like eating and breathing, are stored only with the animal concept. Similarly, properties that are generally true of birds are stored at the bird node, but properties distinctive to individual kinds (e.g., being yellow) are stored with the specific concept nodes they characterize (e.g., CANARY). A property does not have to be true of all subordinate concepts to be stored with a superordinate. This is illustrated in Figure 11.1, where CAN FLY is associated with the bird node; the few exceptions (e.g., flightlessness for ostriches) are stored with particular birds that do not fly. Second, note that category membership is defined in terms of positions in the hierarchical network. For example, the node for CANARY does not directly store the information that canaries are animals; instead, membership would be “computed” by moving from the canary node up to the bird node and then from the bird node to the animal node. It is as if the network presupposes a deductive argument of the form, “All canaries are birds and all birds are animals and therefore all canaries are animals.”

Although these assumptions about cognitive economy and traversing a hierarchical structure may seem speculative, they yield a number of testable predictions. Assuming that traversal takes time, one would predict that the time needed for people to verify properties of concepts should increase with the network distance between the concept and the property. For example, people should be faster to verify that a canary is yellow than to verify that a canary has feathers and faster to determine that a canary can fly than that a canary has skin. Collins and Quillian found general support for these predictions.

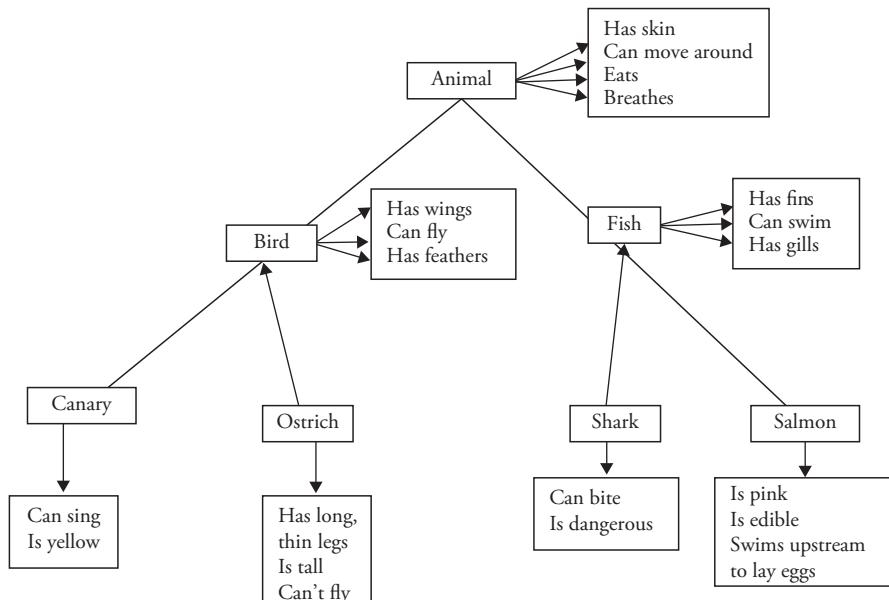


Fig. 11.1 A semantic network after Collins and Quillian (1969).

TURNING POINT 2: NATURAL CONCEPTS AND FAMILY RESEMBLANCE

The work on rule learning suggested that children (and adults) might learn concepts by trying out hypotheses until they hit on the correct definition. In the early 1970s, however, Eleanor Rosch and her associates (e.g., Rosch, 1973; Rosch & Mervis, 1975) argued that most everyday concepts are not organized in terms of the sorts of necessary and sufficient features that would form a (conjunctive) definition for a category. Instead, such concepts depend on properties that are generally true but need not hold for every member. Rosch's proposal was that concepts have a "family-resemblance" structure: What determines category membership is whether an example has enough characteristic properties (is enough like other members) to belong to the category.

One key idea associated with this view is that not all category members are equally "good" examples of a concept. If membership is based on characteristic properties and some members have more of these properties than others, then the ones with more characteristic properties should better exemplify the category. For example, canaries but not penguins have the characteristic bird properties of flying, singing, and building a nest; so one would predict that canaries would be more typical birds than penguins. Rosch and Mervis (1975) found that people do rate some examples of a category to be more typical than

others and that these judgments are highly correlated with the number of characteristic features an example possesses. They also created artificial categories conforming to family-resemblance structures and produced typicality effects on learning and on goodness-of-example judgments.

Rosch and her associates (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976) also argued that the family-resemblance view had important implications for understanding concept hierarchies. Specifically, they suggested that the correlational structure of features (instances that share some features tend to share others) created natural "chunks" or clusters of instances that correspond to what they referred to as *basic-level categories*. For example, having feathers tends to correlate with nesting in trees (among other features) in the animal kingdom, and having gills with living in water. The first cluster tends to isolate birds, while the second picks out fish. The general idea is that these basic-level categories provide the best compromise between maximizing within-category similarity (birds tend to be quite similar to each other) and minimizing between-category similarity (birds tend to be dissimilar to fish). Rosch et al. showed that basic-level categories are preferred by adults in naming objects, are learned first by children, are associated with the fastest categorization reaction times, and have a number of other properties that indicate their special conceptual status.

Turning Points 1 and 2 are not unrelated. To be sure, the Collins and Quillian model, as initially presented, would not predict typicality effects (but see Collins & Loftus, 1975), and it wasn't obvious that it contained anything that would predict the importance of basic-level categories. Nonetheless, these conceptual breakthroughs led to an enormous amount of research premised on the notion that concepts are linked in memory by meaningful pathways, so that memory in effect groups concepts according to their similarity in meaning (see Anderson & Bower, 1973; and Norman & Rumelhart, 1975, for theories and research in this tradition, and Goldstone & Son, Chapter 10, for current theories of similarity).

Fragmentation of Semantics and Memory

Prior to about 1980, most researchers in this field saw themselves as investigating “semantic memory”—the way that long-term memory organizes meaningful information. Around 1980, the term itself became passé, at least for this same group of researchers, and the field regrouped under the banner of “Categories and Concepts” (the title of Smith & Medin’s 1981 synthesis of research in this area). At the time, these researchers may well have seen this change as a purely nominal one, but we suspect it reflected a retreat from the claim that semantic-memory research had much to say about either semantics or memory. How did this change come about?

MEMORY ORGANIZATION

Initial support for a Quillian-type memory organization came from Quillian's own collaboration with Allan Collins (Collins & Quillian, 1969), which we mentioned earlier. Related evidence also came from experiments on lexical priming: Retrieving the meaning of a word made it easier to retrieve the meaning of semantically related words (e.g., Meyer & Schvaneveldt, 1971). In these lexical decision tasks, participants viewed a single string of letters on each trial and decided, under reaction time instructions, whether the string was a word (e.g., “daisy”) or a nonword (“raisy”). The key result was that participants were faster to identify a string as a word if it followed a semantically related item than an unrelated one. For example, reaction time for “daisy” was faster if on the preceding trial the participant had seen “tulip” than if he or she had seen “steel.” This priming effect is consistent with the hypothesis that activation from one concept spreads through memory to semantically related ones.

Later findings suggested, however, that the relation between word meaning and memory organization was less straightforward. For example, the typicality findings (see *Turning Point 2*) suggested that time to verify sentences of the form *An X is a Y* (e.g., “A finch is a bird”) might be a function of the overlap in the information that participants knew about the meaning of *X* and *Y*, rather than the length of the pathway between these concepts. The greater the information overlap—for example, the greater the number of properties that the referents of *X* and *Y* shared—the faster the time to confirm a true sentence and the slower the time to disconfirm a false one. For example, if you know a lot of common information about finches and birds but only a little common information about ostriches and birds, you should be faster to confirm the sentence “A finch is a bird” than “An ostrich is a bird.” Investigators proposed several theories along these lines that made minimal commitments to the way memory organized its mental concepts (McCloskey & Glucksberg, 1979; Smith, Shoben, & Rips, 1974; Tversky, 1977). Rosch’s (1978) theory likewise studiously avoided a stand on memory structure.

Evidence from priming in lexical decision tasks also appeared ambiguous. Although priming occurs between associatively related words (e.g., “bread” and “butter”), it is not so clear that there is priming between semantically linked words in the absence of such associations. It is controversial whether, for example, there is any automatic activation between “glove” and “hat,” despite their joint membership in the clothing category (see Balota, 1994, for a discussion). If memory is organized on a specifically semantic basis—on the basis of word meanings—then there should be activation between semantically related words even in the absence of other sorts of associations. A meta-analysis by Lucas (2000) turned up a small effect of this type, but as Lucas notes, it is difficult to tell whether the semantically related pairs in these experiments are truly free of associations.

The idea that memory organization mimics semantic organization is an attractive one, and memory researchers attempted to modify the original Quillian approach to bring it into line with the results we have just reviewed (e.g., Collins & Loftus, 1975). The data from the sentence verification and lexical decision experiments, however, raised doubts about these theories. Later in this chapter we will consider whether newer techniques can give us a better handle on the structure of memory, but for

now let us turn to the other half of the memory = meaning equation.

SEMANTICS

Specifying the meaning of individual words is one of the goals of semantics, but only one. Semantics must also account for the meaning of phrases, sentences, and longer units of language. One problem in using a theory like Quillian's as a semantic theory is how to extend its core idea—that the meaning of a word is the coordinates of a node in memory structure—to explain how people understand meaningful phrases and sentences. Of course, Quillian's theory and its successors can tell us how we understand sentences that correspond to preexisting memory pathways. We have already seen how the model can explain our ability to confirm sentences like "A daisy is a flower." But what about sentences that do not correspond to preexisting connections, sentences like "Fred placed a daisy in a lunchbox"?

The standard approach to sentence meaning in linguistics is to think of the meaning of sentences as built from the meaning of the words that compose them, guided by the sentence's grammar (e.g., Chierchia & McConnell-Ginet, 1990). We can understand sentences that we have never heard or read before, and since there are an enormous number of such novel sentences, we cannot learn their meaning as single chunks. It therefore seems quite likely that we compute the meaning of these new sentences. But if word meaning is the position of a node in a network, it is hard to see how this position could combine with other positions to produce sentence meanings. What is the process that could take the relative network positions for FRED, PLACE, DAISY, IN, and LUNCHBOX and turn them into a meaning for "Fred placed a daisy in a lunchbox"?

If you like the notion of word meaning as relative position, then one possible solution to the problem of sentence meaning is to connect these positions with further pathways. Since we already have an array of memory nodes and pathways at our disposal, why not add a few more in order to encode the meaning of a new sentence? Perhaps the meaning of "Fred placed a daisy in the lunchbox" is given by a new set of pathways that interconnect the nodes for FRED, PLACE, DAISY, and so on, in a configuration corresponding to the sentence's structure. This is the route that Quillian and his successors took (e.g., Anderson & Bower, 1973; Norman & Rumelhart, 1975; Quillian, 1969), but it comes

at a high price. Adding new connections changes the overall network configuration and thereby alters the meaning of the constituent terms. (Remember: meaning is supposed to be *relative* position.) But it is far from obvious that encoding incidental facts alters word meaning. It seems unlikely, for example, that learning the sentence about Fred changes the meaning of "daisy." Moreover, because meaning is a function of the entire network, the same incidental sentences change the meaning of all words. Learning about Fred's daisy-placing shifts the meaning of seemingly unrelated words like "hippopotamus" if only a bit.

Related questions apply to other psychological theories of meaning in the semantic-memory tradition. To handle the typicality results mentioned earlier, some investigators proposed that the mental representation of a category like daisies consists of a prototype for that category—for example, a description of a good example of a daisy (e.g., Hampton, 1979; McCloskey & Glucksberg, 1979). The meaning of "daisy" in these prototype theories would thus include default characteristics, such as growing in gardens, that apply to most, but not all, daisies. We will discuss prototype theories in more detail soon, but the point for now is that prototype representations for individual words are difficult to combine to obtain a meaning for phrases that contain them. One potential way to combine prototypes—fuzzy set theory (Zadeh, 1965)—proved vulnerable to a range of counterexamples (Osherson & Smith, 1981, 1982). In general, the prototypes of constituent concepts can differ from the prototypes of their combinations in unpredictable ways (Fodor, 1994). The prototype of BIRDS THAT ARE PETS (perhaps a parakeet-like bird) may differ from the prototypes of both BIRDS and PETS (see Storms, de Boeck, van Mechelen, & Ruts, 1998, for related evidence). So if word meanings are prototypes, it is hard to see how the meaning of phrases could be a compositional function of the meaning of their parts.

Other early theories proposed that category representations consist of descriptions of exemplars of the category in question. For example, the mental representation of DAISY would include descriptions of specific daisies that an individual had encoded (e.g., Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1986). However, these theories have semantic difficulties of their own (see Rips, 1995). For example, if, by chance, the only Nebraskans you have met are chiropractors and the

only chiropractors you have met are Nebraskans, then exemplar models appear to mispredict that “Nebraskan” and “chiropractor” will be synonyms for you.

To recap briefly, we have found that experimental research on concepts and categories was largely unable to confirm that global memory organization (as in Quillian's semantic memory) conferred word meaning. In addition, neither the global theories that initiated this research nor the local prototype or exemplar theories that this research produced were able to provide insight into the basic semantic problem of how we understand the meaning of novel sentences. This left semantic-memory theory in the unenviable position of being unable to explain either semantics or memory.

Functions and Findings

Current research in this field still focuses on categorization and communication, but without the benefit of a framework that gives a unified explanation for the functions that concepts play in categorizing, reasoning, learning, language understanding, and memory organization. In this section, we survey the state of the art, and in the following one, we consider the possibility of reuniting some of these roles.

Category Learning and Inference

One nice aspect of Rosch and Mervis's (1975) studies of typicality is that they used both natural language categories and artificially created categories. Finding typicality effects with natural (real-world) categories shows that the phenomenon is of broad interest; finding these same effects with artificial categories provides systematic control for potentially confounding variables (e.g., exemplar frequency) in a way that cannot be done for lexical concepts. This general strategy linking the natural to the artificial has often been followed over the past few decades. Although researchers using artificial categories have sometimes been guilty of treating these categories as ends in themselves, there are enough parallels between results with artificial and natural categories that each area of research informs the other (see Medin & Coley, 1998, for a review).

PROTOTYPE VERSUS EXEMPLAR MODELS

One idea compatible with Rosch's family-resemblance hypothesis is the *prototype view*. It proposes that people learn the characteristic features (or central tendency) of categories and use them

to represent the category (e.g., Reed, 1972). This abstract prototype need not correspond to any experienced example. According to this theory, categorization depends on similarity to the prototypes. For example, to decide whether some animal is a bird or a mammal, a person would compare the (representation of) that animal to both the bird and the mammal prototypes and assign it to the category whose prototype it most resembled. The prototype view accounts for typicality effects in a straightforward manner. Good examples have many characteristic properties of their category and have few characteristics in common with the prototypes of contrasting categories.

Early research appeared to provide striking confirmation of the idea of prototype abstraction. Using random dot patterns as the prototypes, Posner and Keele (1968, 1970) produced a category from each prototype. The instances in a category were “distortions” of the prototype, generated by moving constituent dots varying distances from their original positions. Posner and Keele first trained participants to classify examples that they had created by distorting the prototypes. Then they gave a transfer test in which they presented both the old patterns and new low or high distortions that had not appeared during training. In addition, the prototypes, which the participants had never seen, were presented during transfer. Participants had to categorize these transfer patterns; but unlike the training procedure, the transfer test gave participants no feedback about the correctness of their responses. The tests either immediately followed training or appeared after a 1-week delay.

Posner and Keele (1970) found that correct classification of the new patterns decreased as distortion (distance from a category prototype) increased. This is the standard typicality effect. The most striking result was that a delay differentially affected categorization of prototypic versus old training patterns. Specifically, correct categorization of old patterns decreased over time to a reliably greater extent than performance on prototypes. In the immediate test, participants classified old patterns more accurately than prototypes; but in the delayed test, accuracy on old patterns and prototypes was about the same. This differential forgetting is compatible with the idea that training leaves participants with representations of both training examples and abstracted prototypes, but that memory for examples fades more rapidly than memory for prototypes. The Posner and Keele results were quickly replicated by

others and constituted fairly compelling evidence for the prototype view.

But this proved to be the beginning of the story rather than the end. Other researchers (e.g., Brooks, 1978; Medin & Schaeffer, 1978) put forth an *exemplar view* of categorization. Their idea was that memory for old exemplars by itself could account for transfer patterns without the need for positing memory for prototypes. On this view, new examples are classified by assessing their similarity to stored examples and assigning the new example to the category that has the most similar examples. For instance, some unfamiliar bird (e.g., a heron) might be correctly categorized as a bird not because it is similar to a bird prototype, but rather because it is similar to flamingos, storks, and other shore birds.

In general, similarity to prototypes and similarity to stored examples will tend to be highly correlated (Estes, 1986). Nonetheless, for some category structures and for some specific exemplar and prototype models, it is possible to develop differential predictions. Medin and Schaffer (1978), for example, pitted number of typical features against high similarity to particular training examples and found that categorization was more strongly influenced by the latter. A prototype model would make the opposite prediction.

Another contrast between exemplar and prototype models revolves around sensitivity to within-category correlations (Medin, Altom, Edelson, & Freko, 1982). A prototype representation captures what is on average true of a category but is insensitive to within-category feature distributions. For example, a bird prototype could not represent the impression that small birds are more likely to sing than large birds (unless one had separate prototypes for large and small birds). Medin et al. (1982) found that people are sensitive to within-category correlations (see also Malt & Smith, 1984, for corresponding results with natural object categories). Exemplar theorists were also able to show that exemplar models could readily predict other effects that originally appeared to support prototype theories—differential forgetting of prototypes versus training examples, and prototypes being categorized as accurately or more accurately than training examples. In short, early skirmishes strongly favored exemplar models over prototype models. Parsimony suggested no need to posit prototypes if stored instances could do the job. Since the early 1980s, there have been a number of trends and developments in research and theory with artificially constructed categories,

and we will be able to give only the briefest of summaries here.

NEW MODELS

There are now several contending models for categorizing artificial stimuli, and the early models have been extensively elaborated. For example, researchers have generalized the original Medin and Schaffer (1978) exemplar model to handle continuous dimensions (Nosofsky, 1986), to address the time course of categorization (Lamberts, 1995; Nosofsky & Palmeri, 1997a; Palmeri, 1997), to generate probability estimates in inference tasks (Juslin & Persson, 2002), and to embed it in a neural network (Kruschke, 1992).

Three new kinds of classification theories have been added to the discussion: rational approaches, decision-bound models, and neural network models. Anderson (1990, 1991) proposed that an effective approach to modeling cognition in general and categorization in particular is to analyze the information available to a person in the situation of interest and then to determine abstractly what an effective, if not optimal, strategy might be (see Chater & Oaksford, Chapter 2; Griffiths et al., Chapter 3). This approach has led to some new sorts of experimental evidence (e.g., Anderson & Fincham, 1996; Clapper & Bower, 2002) and pointed researchers more in the direction of the inference function of categories. Interestingly, the Medin and Schaffer exemplar model corresponds to a special case of the rational model, and Nosofsky (1991) has discussed the issue of whether the rational model adds significant explanatory power. However, there is also some evidence undermining the rational model's predictions concerning inference (e.g., Malt, Ross, & Murphy, 1995; Murphy & Ross, 1994; Palmeri, 1999; Ross & Murphy, 1996).

Decision-bound models (e.g. Ashby & Maddox, 1993; Maddox & Ashby, 1993) draw their inspiration from psychophysics and signal detection theory. Their primary claim is that category learning consists of developing decision bounds around the category that will allow people to categorize examples successfully. The closer an item is to the decision bound, the harder it should be to categorize. This framework offers a new perspective on categorization in that it may lead investigators to ask questions such as: How do the decision bounds that humans adopt compare with what is optimal? What kinds of decision functions are easy or hard to acquire? Researchers have also directed efforts

to distinguish decision-bound and exemplar models (e.g., Maddox & Ashby, 1998; Maddox, 1999; McKinley & Nosofsky, 1995; Nosofsky & Palmeri, 1997b; Nosofsky, 1998). One possible difficulty with decision-bound models is that they contain no obvious mechanism by which stimulus familiarity can affect performance, contrary to empirical evidence that it does (Verguts, Storms, & Tuerlinckx, 2003).

Neural network or connectionist models are the third type of new model on the scene (see Knapp & Anderson, 1984, and Kruschke, 1992, for examples; and Doumas & Hummel, Chapter 5, for further discussion of connectionism). It may be a mistake to think of connectionist models as comprising a single category, as they take many forms depending on assumptions about hidden units, attentional processes, recurrence, and the like. There is one sense in which neural network models with hidden units may represent a clear advance on prototype models: They can form prototypes in a bottom-up manner that reflects within-category structure (e.g., Love, Medin, & Gureckis, 2004). That is, if a category comprises two distinct clusters of examples, network models can create a separate hidden unit for each chunk (e.g., large birds versus small birds) and thereby show sensitivity to within-category correlations.

MIXED MODELS AND MULTIPLE CATEGORIZATION SYSTEMS

A common response to hearing about various models of categorization is to suggest that all the models may be capturing important aspects of categorization and that research should determine in which contexts one strategy versus another is likely to dominate. One challenge to this divide-and-conquer program is that the predictions of alternative models tend to be highly correlated and separating them is far from trivial. Nonetheless, there is both empirical research (e.g., Johansen & Palmeri, 2002; Nosofsky, Clark, & Shin, 1989; Reagher & Brooks, 1993) and theoretical modeling that support the idea that mixed models of categorization are useful and perhaps necessary. Current efforts combine rules and examples (e.g., Erickson & Kruschke, 1998; Nosofsky, Palmeri, & McKinley, 1994), as well as rules and decision bounds (Ashby, Alfonso-Reese, Turken, & Waldron, 1998). Some models also combine exemplars and prototypes (e.g., Homa, Sterling, & Trepel, 1981; Minda & Smith, 2001; Smith, Murray, & Minda, 1997; Smith & Minda,

1998, 2000), but it remains controversial whether the addition of prototypes is needed (e.g., Busemeyer, Dewey, & Medin, 1984; Nosofsky & Johansen, 2000; Nosofsky & Zaki, 2002; Stanton, Nosofsky, & Zaki, 2002).

The upsurge of cognitive neuroscience (see Morrison & Knowlton, Chapter 6) has reinforced the interest in multiple memory systems. One intriguing line of research by Knowlton, Squire, and associates (Knowlton, Mangels & Squire, 1996; Knowlton & Squire, 1993; Squire & Knowlton, 1995), favoring multiple categorization systems, involves a dissociation between categorization and recognition. Knowlton and Squire (1993) used the Posner and Keele dot pattern stimuli to test amnesic and matched control patients on either categorization learning and transfer or on a new-old recognition task (involving five previously studied patterns versus five new patterns). The amnesics performed very poorly on the recognition task but were not reliably different from control participants on the categorization task. Knowlton and Squire took this as evidence for a two-system model, one based on explicit memory for examples and one based on an implicit system (possibly prototype abstraction). Reed, Squire, Patalano, Smith, and Jonides (1999) used the same “prototype extraction” task to compare amnesic patients and controls, but instead of dot patterns Reed et al. employed a set of artificial animals that varied on 10 salient features. Unlike the dot patterns, which are quite abstract and difficult to describe, the features of the artificial animals were easy to verbalize. The results, however, were the same: The amnesics performed worse than controls on an explicit-memory test, but they were not reliably different from controls on the categorization task. So implicit category learning is not reserved for what cannot be articulated. On this view, amnesics have lost access to the explicit system but can perform the classification task using their intact implicit memory.

These results have provoked a number of counter-arguments to the hypothesis that categorization and recognition involve two different memory systems. First, investigators have raised questions about the details of the procedures in the prototype-extraction studies. Specifically, Palmeri and Flanery (1999) suggested that some learning may have occurred during the transfer tests (not during training), and this learning did not depend on the storage of exemplars. In support of the claim that learning during transfer occurs in the prototype extraction, Palmeri and Flanery showed that undergraduates who had never

been exposed to training exemplars (the students thought that they were being shown patterns subliminally) performed above chance on transfer tests. Bozoki, Grossman, and Smith (2006) did a similar study, but their subjects included amnesics in addition to normals, and they used artificial animals rather than dot patterns. Bozoki et al. again found that learning occurred even when no training exemplars occurred—for amnesics as well as controls!—but even more learning occurred when exemplars were presented. These results suggest that something is indeed being learned implicitly—namely, whatever it is that both amnesics and controls get from training exemplars. But the two studies under scrutiny differ in too many ways—kinds of materials and subjects—to draw firm conclusions.

Perhaps the most important argument against the Knowlton and Squire evidence for implicit category learning came from Nosofsky and Zaki (1998). They showed that a single-system (exemplar) model could account for both the categorization and recognition data from both the amnesics and controls, by assuming that: (1) the exemplar-based memory of amnesics was impaired but not absent; and (2) the categorization task was less cognitively demanding than the recognition task. A recent review of this literature (Smith, 2008) notes that the Nosofsky-Zaki analysis has never been refuted by patient data (and not for lack of trying).

But there is another kind of data that is relevant here, which turns out to have more diagnostic power—neuroimaging evidence from experiments in which normal subjects perform both prototype-extraction and recognition tasks. Reber, Stark, and Squire (1998a, b) had subjects perform these tasks while having their brains scanned by functional magnetic resonance imaging (fMRI). The recognition task resulted in increased activations in numerous brain areas, including the hippocampus, the very area that is damaged in the amnesiac patients that we have been discussing and that is routinely activated in neuroimaging studies of explicit memory (e.g., Wagner, Bunde, & Bader, 2004). However, none of these areas were activated in the categorization task. Even more striking, an area in visual cortex known to be involved in early visual processing was *activated* in recognition but *deactivated* in categorization. Reber, Gitelman, Parrish, and Mesulam (2003) have observed the same kind of dissociation in another neuroimaging experiment. These results argue quite strongly for a distinction between explicit and implicit category learning.

But what is the nature of the “category” acquired when learning is implicit? The neuroimaging findings suggest an answer. The deactivations observed in occipital cortex are similar to those reported in studies of visual priming (e.g., Buckner, 2000). The latter deactivations are attributed to increased efficiency in processing the perceptual features of the items. This same process could play a role in prototype extraction. The training items are highly similar to one another (else they would not form a category) and thus have numerous features in common. These common features occur frequently during the training phase—else the features would not be “common”—and consequently subjects become increasingly proficient in processing them. If we assume subjects experience a feeling of greater perceptual fluency with items that are processed efficiently, then subjects could use this feeling as a guide to categorization during the transfer phase—the greater the fluency, the more likely the item is a member of the category (Smith, 2008).

So fluency alone can support categorization, but are we really talking about “concepts” here? At the outset, we noted that one of the only things that cognitive scientists agree on concerning concepts is that they are mental representations. Feelings of fluency are not representations of categories (they are more like sensations), so what we have here is categorization without concepts. And perceptual fluency alone may not be able to support the functions that we attributed to concepts early on. Can perceptual fluency support inductive inferences? For example, if one learns a category of artificial animals via perceptual fluency and then learns that some of the instances have an additional feature, will one generalize this feature to other instances? And it seems clear that categories acquired by perceptual fluency will not be available for communication, as it is a hallmark of material learned implicitly that it cannot be verbalized.

The debate is far from resolved, and there are strong advocates both for and against the multiple systems view (e.g., Filoteo, Maddox, & Davis, 2001; Maddox, 2002; Nosofsky & Johansen, 2000; Palmeri & Flanery, 2002; Reber et al., 1998a, b). It is safe to predict that this issue will receive continuing attention.

INFERENCE LEARNING

Recently investigators have begun to worry about extending the scope of category learning by looking at inference. Often we categorize some entity in order to help us accomplish some function or goal.

Ross (1997, 1999, 2000) has shown that the category representations people develop in laboratory studies depend on use and that use affects later categorization. In other words, models of categorization ignore inference and use at their peril. Other work suggests that having a cohesive category structure is more important for inference learning than it is for classification (Yamauchi & Markman, 1998, 2000a, 2000b; Yamauchi, Love, & Markman, 2002; for modeling implications see Love, Markman, & Yamauchi, 2000; Love et al., 2004). More generally, this work raises the possibility that diagnostic rules based on superficial features, which appear so prominently in pure categorization tasks, may not be especially relevant for contexts involving multiple functions or more meaningful stimuli (e.g., Markman & Makin, 1998; Wisniewski & Medin, 1994).

FEATURE LEARNING

The final topic on our “must mention” list for work with artificial categories is feature learning. It is a common assumption in both models of object recognition and category learning that the basic units of analysis or features remain unchanged during learning. There is increasing evidence and supporting computational modeling indicating that this assumption is incorrect. Learning may increase or decrease the distinctiveness of features and may even create new features (see Goldstone, 1998, 2003; Goldstone & Steyvers, 2001; Goldstone, Lippa, & Shriffrin, 2001; Schyns & Rodet, 1997; Schyns, Goldstone, & Thibaut, 1998).

Feature learning has important implications for our understanding of the role of similarity in categorization. It is intuitively compelling to think of similarity as a causal factor supporting categorization—things belong to the same category because they are similar. But this may have things backward. Even standard models of categorization assume that learners selectively attend to features that are diagnostic, and the work on feature learning suggests that learners may create new features that help partition examples into categories. In that sense, similarity (in the sense of overlap in features) is the byproduct, not the cause, of category learning. We will take up this point in discussing the theory of categorization later in this review.

FEATURES OF NATURAL CATEGORIES AND COGNITIVE NEUROSCIENCE

In the research reviewed thus far, the features (1) have not been subdivided into types and (2) have

been assumed to be abstract symbols. Both of these assumptions have been challenged by the foray of cognitive neuroscience into the study of natural concepts, a movement that began in the mid-1980s but still remains somewhat separate from the mainstream of behavioral studies on categories and concepts. In what follows, we try to convey the gist of this research by focusing on studies dealing with Points (1) and (2) mentioned earlier. This work includes both neuropsychological studies of neurological patients with categorization deficits and neuroimaging studies of normals (typically college students) who answer questions about categories and their properties while having their brains scanned.

Neuropsychological Studies of Categorization

The seminal article in this line of research was a 1984 paper by Warrington and Shallice that analyzed the categorization abilities of patients with “visual agnosia,” an inability to categorize visual objects even though perceptual abilities remain intact. Such patients often have lesions in the bottom or ventral portion of temporal cortex. It had been known for some time that such patients may show greater deficits with some categories than others. Warrington and Shallice (1984) thoroughly documented such “category-specific” deficits in four patients, showing that patients could have normal levels of categorization accuracy on some categories while being drastically impaired on others. For example, in a task in which subjects simply had to match pictures to category-names, Patient JBR achieved 98% accuracy on categories of nonliving things (obviously normal), but only 67% on categories of animals (obviously impaired). Other patients showed a similar pattern. However, in some tasks there were exceptions to the living versus nonliving distinction, such as musical instruments, which clustered with living things, whereas body parts clustered with the nonliving categories.

To handle these exceptions, Warrington and Shallice (1984) shifted their analysis from the level of categories to that of features. They noted that the impaired categories tended to be those characterized by perceptual features, while the less impaired, or intact, categories tended to be characterized by functional features (where “functional” is used very broadly). This led to the “perceptual-functional” hypothesis of category deficits: Brain damage can selectively affect sensory/perceptual regions or regions more involved in the use of artifacts

(e.g., motor areas), and the first type of damage will hurt the categorization of living things while the second will mainly affect categories of nonliving things like tools. The hypothesis is extremely influential in neurological and neuroimaging studies, and it makes clear the importance of distinguishing features with regard to what they refer to—perceptual versus functional properties of objects.

If problems with perceptual features are what underlie patients' problems in dealing with animal categories, one would expect that, when asked about the features of a category (e.g., "Does a canary have a beak?") these patients should have more problems with questions about perceptual than nonperceptual features of the same category. This prediction has been confirmed numerous times (e.g., Thompson-Schill, D'Esposito, Aguirre, & Farah, 1997). The difference between perceptual and functional features seems to be a key distinction about the nature of concepts that has been relatively ignored in the mainstream literature on normal categorization.

This perceptual-functional distinction lies at the heart of a number of computational models of categorization that have been applied to the results of the neuropsychological studies. Farah and McClelland (1991) developed a model in which conceptual knowledge about objects is divided into perceptual and functional components, and the ratio of perceptual-to-functional features for living things is much greater than that for nonliving things. They then selectively degraded either perceptual or functional features and showed that this "lesioned" model provided a good account of patients' performance on simple categorization tasks. Subsequent models have become increasingly more sophisticated. Thus, in Rogers and McClelland's (2004) model, the features are not built into the model by the modeler, but rather are based on normal subjects' verbal descriptions and drawings of these objects. These empirically obtained features reveal not only the perceptual-functional distinction but also a further distinction based on whether color is a feature, which is critical in distinguishing animals from plants (fruits/vegetables), both of which are living things. When the different kinds of features are "lesioned," the model's performance on a variety of simple categorization tasks matches that of various patients to a striking degree.

Neuroimaging Studies

The first neuroimaging studies of natural concepts appeared in the mid-1990s. Martin, Haxby,

Lalonde, Wiggs, and Ungerleider (1995) imaged peoples' brains while they performed a simple property-verification task (e.g., "Is a banana yellow?"). Unsurprisingly they found increased activation in left-hemisphere language areas (located in the ventral part of the frontal cortex and the medial part of the temporal cortex). In addition, though, there was a notable increase in activation in the ventral region of temporal cortex, which is the region responsible for object recognition. Even though only semantic processing was called for, object-processing seemed to be involved as well. Even more striking, the activation at issue was close to the subregion that underlies the perception of the property. For example, when answering verbal questions about the color of an object, there was an activation in the area concerned with perceiving the color of objects.

Subsequent studies replicated the critical finding that perceptual questions activate perceptual areas—what has come to be called "modality specificity" (e.g., Chao & Martin, 1999; Kellenbach, Brett, & Patterson, 2001). Indeed, some studies showed activation in perceptual areas even when language areas were not activated (Kan, Barasalou, Solomon, Minor, & Thompson-Schill, 2003). The neuroimaging evidence for modality specificity gets even more detailed, as some property-verification studies have varied the kind of perceptual property queried—color versus size, for example—and shown that color questions activate color areas, whereas size questions activate areas in parietal cortex known to be involved in the perception of space. Interestingly, though the brain areas being activated are clearly perceptual ones, subjects in these experiments rarely if ever report experiencing visual imagery. The conclusion that has repeatedly been drawn from these studies is that knowledge about perceptual properties, which had been thought to be represented in a completely amodal fashion, is in fact partially represented in a modality-specific way (Thompson-Schill, 2003). Object concepts are claimed to be "perceptually embodied." Furthermore, since the different perceptual modalities are processed in different brain regions, "perceptual-conceptual knowledge" must be represented by a distributed network of brain regions. There is no Bird area in the brain.

Modality specificity cannot be the whole story, however. There are longstanding arguments against the idea that object concepts are purely perceptual (summarized in Miller & Johnson-Laird, 1976). For example, why do people who are congenitally blind seem to share the same concepts as the

sighted? It is even possible to question whether the modality-specific activations found in the aforementioned studies actually play an important role in categorization and property verification (e.g., Mahon & Caramazza, 2008). Specifically, the knowledge that is used to answer simple questions about object categories may be represented amodally (symbolically), as has routinely been assumed in the semantic-memory literature, but these representations may be closely linked to modality-specific ones, which may be used for purposes of recognizing objects and generating visual images. If so, activation of the relevant amodal representation may lead to the concurrent activation of the associated modality-specific representation. The apparent evidence for modality specificity is essentially a spread of activation from the amodal areas that do the real neural work. At this point, however, there is little independent evidence in favor of the spreading activation account, so modality specificity cannot be dismissed.

REASONING

As we noted earlier, one of the central functions of categorization is to support reasoning. Having categorized some entity as a bird, one may predict with reasonable confidence that it builds a nest, sings, and can fly, though none of these inferences is certain. In addition, between-category relations may guide reasoning. For example, from the knowledge that robins have some enzyme in their blood one is likely to be more confident that the enzyme is in sparrows than in raccoons. The basis for this confidence may be that robins are more similar to sparrows than to raccoons, or that robins and sparrows are more likely to share relevant biological mechanisms than robins and raccoons. We will not review this literature here, but see Heit (2000), Kemp and Tenenbaum (2009), and Rips (2001, 2011) for detailed treatments.

SUMMARY

Due to space limitations, we have glossed a lot of research and skipped numerous other relevant studies. The distinction between artificially created and natural categories is itself artificial, at least in the sense that it has no clear definition or marker. When we take up the idea that concepts may be organized in terms of theories, we will return to some laboratory studies that illustrate this fuzzy boundary. For the moment, however, we shift attention to the more language-like functions of concepts.

Language Functions

Most investigators in the concepts-and-categories area continue to assume that, in addition to their role in recognition and category learning, concepts also play a role in understanding language and in thinking discursively about things. In addition to determining, for example, which perceptual patterns signal the appearance of a daisy, the DAISY concept also contributes to the meaning of sentences like our earlier example, “Fred placed a daisy in a lunchbox.” We noted that early psychological research on concepts ran into problems in explaining the meaning of linguistic units larger than single words. Most early theories posited representations, such as networks, exemplars, or prototypes, that did not combine easily and thus complicated the problem of sentence meaning. Even if we reject the idea that sentence meanings are compositional functions of word meaning, we still need a theory of sentence meanings, and no obvious contenders are in sight. In this section, we return to the role that concepts play in language to see whether new experiments and theories have clarified this relationship.

CONCEPTS AS POSITIONS IN MEMORY STRUCTURES

One difficulty with the older semantic-memory view of word meaning is that memory seems to change with experience from one person to another, while meaning must be more or less constant. The sentences that you have encoded about daisies may differ drastically from those that we have encoded, because your conversation, reading habits, and other verbal give-and-take can diverge in important ways from ours. If meaning depends on memory for these sentences, then your meaning for “daisy” should likewise differ from ours. This observation raises the question of how you could possibly understand the sentences in this chapter in the way we intend or how you could meaningfully disagree with us about some common topic (see Fodor, 1994).

Two people—say, Calvin and Martha—might be able to maintain mutual intelligibility as long as their conceptual networks are not too different. It is partly an empirical question how much their networks can vary while still allowing Calvin’s concepts to map correctly into Martha’s. To investigate this issue, Goldstone and Rogosky (2002) carried out some simulations that try to recover such a mapping. The simulations modeled Calvin’s conceptual system as the distance between each pair of his concepts (e.g., the distance between DOG and

CAT in Calvin's system might be one unit, while the distance between DOG and DAISY might be six units). Martha's conceptual system was represented in the same way (i.e., by exactly the same interconcept distances), except for random noise that Goldstone and Rogosky added to each distance to simulate the effect of disparate beliefs. A constraint-satisfaction algorithm was then applied to Calvin's and Martha's systems that attempted to recover the original correspondence between the concepts—to map Calvin's DOG to Martha's DOG, Calvin's DAISY to Martha's DAISY, and so on. The results of the simulations show that with 15 concepts in each system (the maximum number considered and the case in which the model performed best) and with no noise added to Martha's system, the algorithm was always able to find the correct correspondence. When the simulation added to each dimension of the interconcept distance in Martha a small random increment (drawn from a normal distribution with mean 0 and standard deviation equal to .004 times the maximum distance), the algorithm recovered the correspondence about 63% of the time. When the standard deviation increased to .006 times the maximum distance, the algorithm succeeded about 15% of the time (Goldstone & Rogosky, 2002, Figure 2).

What should one make of the Goldstone and Rogosky results? Correspondences may be recovered for small amounts of noise, but performance dropped off dramatically for larger amounts of noise. Foes of the meaning-as-relative-position theory might claim that the poor performance under 0.6% noise proves their contention. Advocates would point to the successful part of the simulations and note that its ability to detect correct correspondences usually improved as the number of points increased (although there are some non-monotonocities in the simulation results that qualify this finding). Clearly, this is only the beginning of the empirical side of the debate. For example, the differences between Martha and Calvin are likely to be not only random but also systematic, as in the case where Martha grew up on a farm and Calvin was a city kid.

CONCEPT COMBINATION

Let us now consider attempts to tackle head-on the problem of how word-level concepts combine to produce the meanings of larger linguistic units. There is relatively little research in this tradition on entire sentences (see Conrad & Rips, 1986; Rips,

Smith, & Shoben, 1978), but there has been a fairly steady research stream devoted to noun phrases, including adjective-noun ("edible flowers"), noun-noun ("food flowers"), and noun-relative-clause combinations ("flowers that are foods"). We will call the noun or adjective parts of these phrases *components* and distinguish the main or *head noun* ("flowers" in each of our examples) from the adjective or noun *modifier* ("edible" or "food"). The aim of the research in question is to describe how people understand these phrases and, in particular, how the typicality of an instance in these combinations depends on the typicality of the same instance in the components. How does the typicality of a marigold in the category of edible flowers depend on the typicality of marigolds in the categories of edible things and flowers? As we have already noticed, this relationship is far from straightforward (parakeets are superbly typical as pet birds but less typical pets and even less typical birds). For reviews of work on this problem, see Hampton (1997), Murphy (2002), Rips (1995), and Wisniewski (1997).

Research on the typicality structure of combinations has turned up interesting phenomena. These same phenomena, however, pose some serious challenges to the idea that people can predict the typicality of a combination from the typicality of its components. This difficulty is instructive, in part because all psychological theories of concept combination posit complex, structured representations, and they depict concept combination either as rearranging (or augmenting) the structure of the head noun by means of the modifier (Franks, 1995; Murphy, 1988; Smith, Osherson, Rips, & Keane, 1988) or as fitting both head and modifier into a larger relational complex (Gagné & Shoben, 1997). Table 11.1 summarizes the assumptions of these theories. Earlier models (at the top of the table) differ from later ones mainly in terms of the complexity of the combination process. Smith et al. (1988), for example, aimed at explaining simple adjective-noun combinations (e.g., "white vegetable") that, roughly speaking, refer to the intersection of the sets denoted by modifier and head (white vegetables are approximately the intersection of white things and vegetables). In this theory, combination occurs when the modifier changes the value of an attribute in the head noun (changing the value of the color attribute in VEGETABLE to WHITE) and boosts the importance of this attribute in the overall representation. Later theories attempted to account for nonintersective combinations (e.g., "criminal

Table 11.1. Some Theories of Concept Combination

Model	Domain	Representation of Head Noun	Modification Process
Hampton (1987)	Noun-Noun and Noun-Relative-Clause NPs (conjunctive NPs, e.g., <i>sports that are also games</i>)	Schemas (attribute-value lists with attributes varying in importance)	Modifier and head contribute values to combination on the basis of importance and centrality
Smith, Osherson, Rips, & Keane (1988)	Simple Adjective-Noun NPs (e.g., <i>red apple</i>)	Schemas (attribute-value lists with distributions of values and weighted attributes)	Adjective shifts value on relevant attribute in head and increases weight on relevant dimension.
Murphy (1988)	Adj-Noun and Noun-Noun NPs (esp. non-predicating NPs, e.g., <i>corporate lawyer</i>)	Schemas (lists of slots and fillers)	Modifier fills relevant slot; then representation is “cleaned up” on the basis of world knowledge.
Franks (1995)	Adj-Noun and Noun-Noun NPs (esp. privatives, e.g., <i>fake gun</i>)	Schemas (attribute-value structures with default values for some attributes)	Attribute-values of modifier and head are summed, with modifier potentially overriding or negating head values.
Gagné & Shoben (1997)	Noun-Noun NPs	Lexical representations containing distributions of relations in which nouns figure	Nouns are bound as arguments to relations (e.g., <i>flu virus</i> = virus causing flu).
Wisniewski (1997)	Noun-Noun NPs	Schemas (lists of slots and fillers, including roles in relevant events)	<ol style="list-style-type: none"> 1. Modifier noun is bound to role in head noun (e.g., <i>truck soap</i> = soap for cleaning trucks). 2. Modifier value is reconstructed in head noun (e.g., <i>zebra clam</i> = clam with stripes). 3. Hybridization (e.g., <i>robin canary</i> = cross between robin and canary)

lawyers,” who are often not both criminals and lawyers). These combinations call for more complicated adjustments, for example, determining a relation that links the modifier and head (a criminal lawyer is a lawyer whose clients are in for criminal charges) or extracting a value from the modifier that can then be assigned to the head (e.g., a panther lawyer might be one who is especially vicious or tenacious).

One reason for difficulty in providing a theory for concept combination is that many of the combinations that investigators have studied are familiar or, at least, have familiar referents. Some people have experience with edible flowers, for example,

and know that they include nasturtiums, are sometimes used in salads, are often brightly colored, are peppery-tasting, and so on. We learn many of these properties by direct or indirect observation (by what Hampton, 1987, calls “extensional feedback”), and they are sometimes impossible to learn simply by knowing the meaning of “edible” and “flower.” Because these properties can affect the typicality of potential instances, the typicality of these familiar combinations won’t be a function of the typicality of their components. This means that if we are going to be able to predict typicality in a compositional way, we will have to factor out the contribution of these directly acquired properties. Rips (1995) refers

to this filtering as the “No Peeking Principle”—no peeking at the referents of the combination. Of course, you might be able to predict typicality if you already know the relevant real-world facts in addition to knowing the meaning of the component concepts. The issue about understanding phrases, however, is how we are able to interpret an unlimited number of new ones. For this purpose, people need some procedure for computing new meanings from old ones that is not restricted by the limited set of facts they happened to have learned (e.g., through idiosyncratic encounters with edible flowers).

Another reason why it is hard to explain concept combination is that some of the combinations used in this research may be compounds or lexicalized phrases (e.g., “White House” [accent on “White”] = the residence of the President) rather than modifier-head constructions (e.g., “white house” [accent on “house”] = a house whose color is white). Compounds are often idiomatic; their meaning is not an obvious function of their parts (see Gleitman & Gleitman’s, 1970, distinction between phrasal and compound constructions; and Partee, 1995).

There is a deeper reason, however, for the difficulty in predicting compound typicality from component typicality. Even if we adhere to the No Peeking Principle, and even if we stick to clear modifier-head constructions, the typicality of a combination can depend on “emergent” properties that are not part of the representation of either component (Hastie, Schroeder, & Weber, 1990; Kunda, Miller, & Claire, 1990; Medin & Shoben, 1988; Murphy, 1988). For example, you may never have encountered, or even thought about, a smoky apple (so extensional feedback does not inform your conception of the noun phrase), but nevertheless it is plausible to suppose that smoky apples are not good tasting. Having a bad taste, however, is not a usual property of (and is not likely to be stored as part of a concept for) either apples or smoky things; on the contrary, many apples and smoky things (e.g., smoked meats, cheeses, and fish) are quite good tasting. If you agree with our assessment that smoky apples are likely to be bad tasting, that is probably because you imagine a way in which apples could become smoky (being caught in a kitchen fire, perhaps) and you infer that under these circumstances the apple would not be good to eat. The upshot is that the properties of a combination can depend on complex inductive or explanatory inferences (Johnson & Keil, 2000; Kunda et al., 1990). If these properties affect the typicality of an instance with

respect to the combination, then models of this phenomenon will not be simple. No current theory provides an adequate and general account of these processes.

INFERENTIAL VERSUS ATOMISTIC CONCEPTS

Research on the typicality structure of noun phrases is of interest for what it can tell us about people’s inference and problem-solving skills. But because these processes can be quite complex—drawing on general knowledge and inductive reasoning to produce emergent information—we cannot predict noun phrase typicality in other than a limited range of cases. By themselves, emergent properties do not rule out the possibility of a model that explains how people derive the meaning of a noun phrase from the meaning of its components. Compositionality does not require that all aspects of the noun phrase’s meaning are literally parts of the components’ meanings. It is sufficient to find some computable function from the components to the composite that is simple enough to account for people’s understanding (see Partee, 1995, for a discussion of types of composition). The trouble is that if noun phrases’ meanings require theory construction and problem solving, such a process is unlikely to explain the ease and speed with which we usually understand them in ongoing speech.

We have only considered the role of networks or prototypes in concept combination, but many of the same difficulties with semantic composition affect other contemporary theories, such as latent semantic analysis (Landauer & Dumais, 1997), which take a global approach to meaning. Latent semantic analysis takes as input a table of the frequencies with which words appear in specific contexts. In one application, for example, the items comprise about 60,000 word types taken from 30,000 encyclopedia entries, and the table indicates the frequency with which each word appears in each entry. The analysis then applies a technique similar to factor analysis to derive an approximately 300-dimensional space in which each word appears as a point and in which words that tend to co-occur in context occupy neighboring regions in the space. Because this technique finds a best fit to a large corpus of data, it is sensitive to indirect connections between words that inform their meaning. However, the theory has no clear way to derive the meaning of novel sentences. Although latent semantic analysis could represent a sentence as the average position of its component words, this would not allow it to

capture the difference between, say, *The financier dazzled the movie star* versus *The movie star dazzled the financier*, which depend on sentence structure as well as word meaning. In addition, the theory uses the distance between two words in semantic space to represent the relation between them; so the theory has trouble with semantic relations that, unlike distances, are asymmetric. It is unclear, for example, how it could cope with the fact that *father* implies *parent* but *parent* does not imply *father*.

On the one hand, online sentence understanding is a rapid, reliable process. On the other, the meaning of even simple adjective-noun phrases seems to require heady inductive inferences. Perhaps we should distinguish, then, between the *interpretation* of a phrase or sentence and its *comprehension* (Burge, 1999). On this view, comprehension gives us a more or less immediate understanding of novel phrases, based primarily on the word meaning of the components and syntactic/semantic structure. Interpretation, by contrast, is a potentially open-ended process, relying on the result of comprehension plus inference and general knowledge. As we have just noticed, it is hard, if not impossible, to compute the typicality structure of composites. So if we want something readily computable in order to account for comprehension, we have to look to something simpler than typicality structures (and the networks, prototypes, schemas, or theories that underlie them). One possibility (Fodor, 1994, 1998) is to consider a representation in which word meanings are mental units not much different from the words themselves, and whose semantic values derive from (unrepresented) causal connections to their referents.

The comprehension/interpretation distinction that we just proposed may be more of a continuum than a dichotomy. Background knowledge about the component concepts may come in gradually as part of sentence interpretation, with more atomistic components becoming available early in the process, simple typicality information arriving later, and inference-intensive information still later. Researchers usually study sentence understanding, classification, and inference with different methodologies. Reaction time, for example, is often a dependent measure for sentence understanding and classification, but rarely for interpretation tasks, which employ off-line choice and think-aloud methods (more on interpretation tasks in upcoming sections). Different methodologies make it hard to generalize about the way in which knowledge

informs sentence understanding. However, some evidence exists that typicality information, in particular, can impact simple language generation tasks. In early work, Rosch (1975) asked one group of subjects to produce sentences containing names of categories like “bird” and “weapon.” The result was often items like “Twenty or so birds often perch on the telephone wires outside my window and twitter in the morning.” Rosch substituted for the category name in these sentences either the name of a typical member of the category (“sparrow”) or an atypical member (“turkey”). An independent group of subjects then judged how natural or peculiar the resulting sentences were, and they gave higher naturalness ratings to the sentences with typical instances (“Twenty or so robins often perch on the telephone wires...”) than sentences with atypical instances (“Twenty or so turkeys often perch on the telephone wires...”). Typicality therefore seems to inform simple sentence production.

Returning to concept combination, we might be able to salvage some of the models in Table 11.1 if we view the models as describing a particular stage of the comprehension process. Theories like Smith et al.’s (1988) Selective Modification Model, for example, might give the right account of concept combination for a stage that is not too early (so some typicality information can enter the picture) and not too late (so higher level inferences do not have time to override typicality).

GENERIC NOUN PHRASES

Even if we abandon typicality structures as accounts of comprehension, however, it does not follow that these structures are useless in explaining all linguistic phenomena. Recent research on two fronts seems to us to hold promise for interactions between psychological and linguistic theories. First, there are special constructions in English that, roughly speaking, describe default characteristics of members of a category. For example, “Lions have manes” means (approximately) that having a mane is a characteristic property of lions. Bare plural noun phrases (i.e., plurals with no preceding determiners) are one way to convey such a meaning as we have just seen, but indefinite singular sentences (“A lion has a mane”) and definite singular sentences (“The lion—*Panthera leo*—has a mane”) can also convey the same idea in some of their senses. These generic sentences seem to have normative content. Unlike “Most lions have manes,” generic sentences seem to hold despite the existence of numerous exceptions;

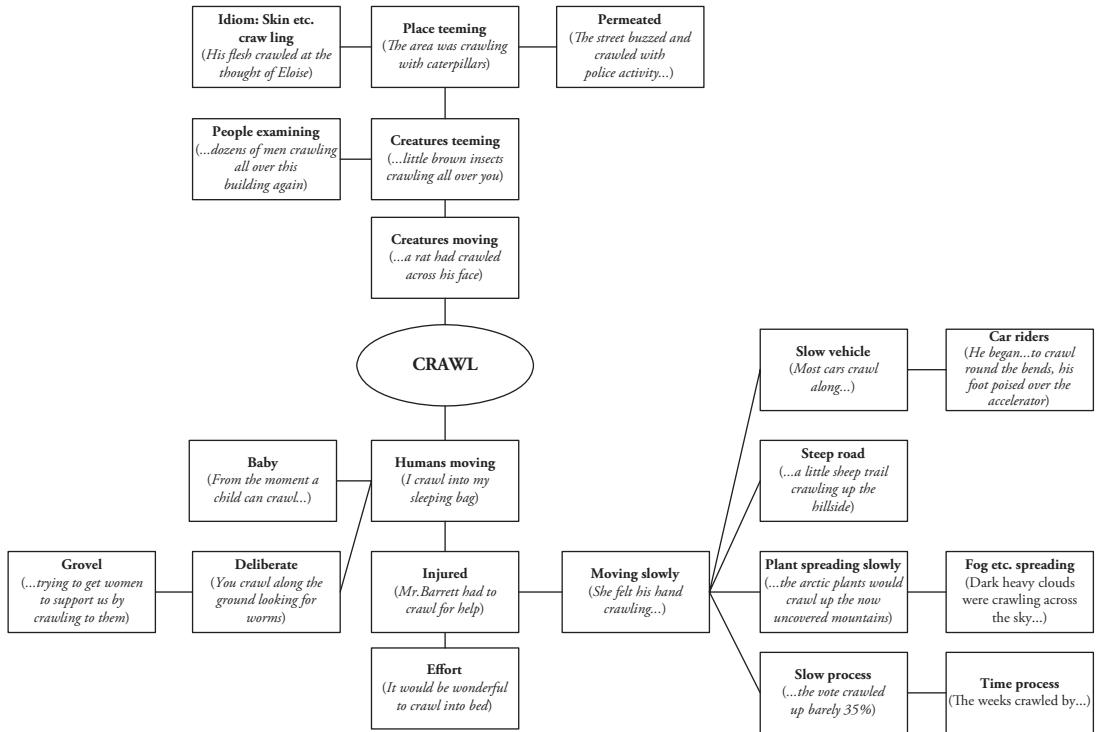


Fig. 11.2 The meaning of crawl: Why it is difficult to distinguish different related meanings (polysemy) from different uses of the same meaning (contextual variation) and from different unrelated meanings (homonymy). Adapted from Fillmore & Alking (2000) by permission of Oxford University Press.

“Lions have manes” seems to be true despite the fact that most lions (e.g., female and immature lions) do not have manes (see Krifka et al., 1995, for an introduction to generic sentences, and Leslie, 2008, for one recent account). There is an obvious relation between the truth or acceptability of generic sentences and the typicality structure of categories, since the typical properties of a category are those that appear in true generic sentences. Of course, as Krifka et al. note, this may simply be substituting one puzzle (the truth conditions of generic sentences) for another (the nature of typical properties), but this may be one place where linguistic and cognitive theories might provide mutual insight. Research by Susan Gelman and her colleagues (see Gelman, 2003, for a thorough review) suggests that generic sentences are a frequent way for caregivers to convey category information to children. Four-year-olds differentiate sentences with bare plurals (“Lions have manes”) from those explicitly quantified by “all” or “some” in comprehension, production, and inference tasks (Gelman, Star, & Flukes, 2002; Hollander, Gelman, & Star, 2002). It would be of interest to know, however, exactly

what meaning children and adults attach to generic sentences.

POLYSEMY

A second place to look for linguistic-cognitive synergy is in an account of the meanings of polysemous words. Linguists (e.g., Lyons, 1977) traditionally distinguish homonyms like “mold,” which have multiple unrelated meanings (e.g., a form into which liquids are poured vs. a fungus) from polysemous terms like “line” that have multiple related meanings (e.g., a geometric line vs. a fishing line vs. a line of people, etc.). What makes polysemous terms interesting to psychologists in this area is that the relations among their meanings often possess a kind of typicality structure of their own. This is the typicality of the senses of the expression rather than the typicality of the referents of the expression, making it a kind of higher level typicality phenomenon. Figure 11.2 illustrates such a structure for the polysemous verb “crawl,” as analyzed by Fillmore and Atkins (2000). A rectangle in the figure represents each sense or use and includes both a brief label indicating its distinctive property and an example

from British corporuses. According to Fillmore and Atkins, the central meanings for *crawl* have to do with people or creatures moving close to the ground (these uses appear in rectangles with darker outlines in the figure). But there are many peripheral uses, for example, time moving slowly ("The hours seemed to crawl by") and creatures teeming about ("The picnic supplies crawled with ants"). The central meanings are presumably the original ones, with the peripheral meanings derived from these by a historical chaining process. Malt, Sloman, Gennari, Shi, and Wang (1999) have observed similar instances of chaining in people's naming of artifacts, such as bottles and bowls, and it is possible that the gerrymandered naming patterns reflect the polysemy of the terms (e.g., "bottle") rather than different uses of the same meaning. As Figure 11.2 shows, it is far from easy to distinguish different related meanings (polysemy) from different uses of the same meaning (contextual variation) and from different unrelated meanings (homonymy).

Some research has attacked the issue of whether people store each of the separate senses of a polysemous term (Klein & Murphy, 2002) or store only the core meaning, deriving the remaining senses as needed for comprehension (Caramazza & Grober, 1976; Franks, 1995). Conflicting evidence in this respect may be due to the fact that some relations between senses seem relatively productive and derivable (*regular* polysemy), such as the relationship between terms for animals and their food products, e.g., the animal meaning of "lamb" and its menu meaning), while other senses seem ad hoc (e.g., the relation between "crawl" = *moving close to the ground* and "crawl" = *teeming with people* in Fig. 11.2). Multiple mechanisms are likely to be at work here.

SUMMARY

We do not mean to suggest that the only linguistic applications of psychologists' "concepts" are in dealing with interpretation, generic phrases, and polysemy. Far from it. There are many areas, especially in developmental psycholinguistics, that hold the promise of fruitful interactions but that we cannot review here. Nor are we suggesting that investigators in this area give up the attempt to study the use of concepts in immediate comprehension. But concepts for comprehension seem to have different properties from the concepts that figure in the other functions we have discussed, and researchers need to direct more attention to the interface between them.

Theories, Modules, and Psychometaphysics

We have seen, so far, some downward pressure on cognitive theories to portray human concepts as mental entities that are as simple and streamlined as possible. This pressure comes not only from the usual goal of parsimony but also from the role that concepts play in immediate language comprehension. But there is also a great deal of upward pressure, pressure to include general knowledge about a category as part of its representation. For example, the presence of emergent properties in concept combinations suggests that people use background knowledge in interpreting these phrases. Similarly, people may bring background knowledge and theories to bear in classifying things, even when they know a decision rule for the category. Consider psychodiagnostic classification. Although *DSM-IV* (the official diagnostic manual of the American Psychological Association) is atheoretical and organized in terms of rules, there is clear evidence that clinicians develop theories of disorders and, contra *DSM-IV*, weight causally central symptoms more than causally peripheral symptoms (e.g., Kim & Ahn, 2002a). The same holds for laypeople (e.g. Furnham, 1995; Kim & Ahn, 2002b).

In this section we examine the consequences of expanding the notion of a concept to include theoretical information about a category. In the case of the natural categories, this information is likely to be causal (see Buehner & Cheng, Chapter 12), since people probably view physical causes as shaping and maintaining these categories. For artifacts, the relevant information may be the intentions of the person creating the object (e.g., Bloom, 1996). The issues we raise here concern the content and packaging of these causal beliefs.

The first of these issues focuses on people's beliefs about the locus of these causal forces—what we called "psychometaphysics." At one extreme, people may believe that each natural category is associated with a single source, concentrated within a category instance, that controls the nature of that instance. The source could determine, among other things, the instance's typical properties, its category membership, and perhaps even the conditions under which it comes into and goes out of existence. Alternatively, people may believe that the relevant causal forces are more like a swarm—not necessarily internal to an instance, nor necessarily emanating from a unitary spot—but shaping the category in aggregate fashion.

The second issue has to do with the cognitive divisions that separate beliefs about different sorts

of categories. People surely think that the causes that help shape daisies differ in type from those that shape teapots. Lay theories about flowers and other living things include at least crude information about specifically biological properties, whereas lay theories of teapots and other artifacts touch instead on intended and actual functions. But how deep do these divisions go? On the one hand, beliefs about these domains could be modular (relatively clustered, relatively isolated), innate, universal, and local to specific brain regions. On the other, they may be free-floating, learned, culturally specific, and distributed across cortical space. This issue is important to us because it ultimately affects whether we can patch up the “semantic memory” marriage.

Essentialism and Sortalism

PSYCHOLOGICAL ESSENTIALISM

What is the nature of people’s beliefs about the causes of natural kinds? One hypothesis is that people think there is something internal to each member of the kind—an essence—that is responsible for its existence, category membership, typical properties, and other important characteristics (e.g., Atran, 1998; Gelman & Hirschfeld, 1999; Medin & Ortony, 1989). Of course, it is unlikely that people think that all categories of natural objects have a corresponding essence. There is probably no essence of pets, for example, that determines an animal’s pet status. But for basic-level categories, such as dogs or gold or daisies, it is tempting to think that something in the instance determines crucial aspects of its identity. Investigators who have accepted this hypothesis are quick to point out that the theory applies to people’s beliefs and not to the natural kinds themselves. Many biologists and philosophers of science agree that essentialism will not account for the properties and variations that real species display, in part because the very notion of species is not coherent (e.g., Ghiselin, 1981; Hull, 1999). Chemical kinds (e.g., gold) may conform much more closely to essentialist doctrine (see Sober, 1980). Nevertheless, expert opinion is no bar to laypersons’ essentialist views on this topic. In addition, psychological essentialists have argued that people probably do not have a fully fleshed out explanation of what the essence is. What they have, on this hypothesis, is an IOU for a theory: a belief that there must be *something* that plays the role of essence, even though they cannot supply a description of it (Medin & Ortony, 1989).

Belief in a hypothetical, minimally described essence may not seem like the sort of thing that could do important cognitive work, but psychological essentialists have pointed out a number of advantages that essences might afford, especially to children. The principle advantage may be induction potential. Medin (1989) suggested that essentialism is poor metaphysics but good epistemology in that it may lead people to expect that members of a kind will share numerous, unknown properties—an assumption that is sometimes correct. In short, essences have a motivational role to play in getting people to investigate kinds’ deeper characteristics. Essences also explain why category instances seem to run true to type—for example, why the offspring of pigs grow up to be pigs rather than cows. They also explain the normative character of kinds (e.g., their ability to support inductive arguments and their ability to withstand exceptions and superficial changes), as well as people’s tendency to view terms for kinds as well defined.

Evidence for essentialism tends to be indirect. There are results that show that children and adults do in fact hold the sorts of beliefs that essences can explain. By the time they reach first or second grade, children know that animals whose insides have been removed are no longer animals, that baby pigs raised by cows grow up to be pigs rather than cows (Gelman & Wellman, 1991), and that cosmetic surgery does not alter basic-level category membership (Keil, 1989). Research on adults also shows that “deeper” causes—those that themselves have few causes but many effects—tend to be more important in classifying than shallower causes (Ahn, 1998; Sloman, Love, & Ahn, 1998; but see Rehder, 2009).

However, results like these are evidence for essence only if there are no better explanations for the same results, and it seems at least conceivable that children and adults make room for multiple types and sources of causes that are not yoked to an essence. According to Strevens (2000), for example, although people’s reasoning and classifying suggest that causal laws govern natural kinds, it may be these laws alone, rather than a unifying essence, that are responsible for the findings. According to essentialists, people think there is something (an essence) that is directly or indirectly responsible for the typical properties of a natural kind. According to Strevens’ minimalist alternative, people think that for each typical property there is something that causes it, and that something may vary for different properties.

Essentialists counter that both children and adults assume a causal structure consistent with essence (see Gelman & Frazier, Chapter 26; also see Braisby, Franks, & Hampton, 1996; Diesendruck & Gelman, 1999; Kalish, 1995, 2002, for debate on this issue). One strong piece of evidence for essentialism is that participants who have successfully learned artificial, family-resemblance categories (i.e., those in which category members have no single feature in common) nevertheless believe that each category contained a common, defining property (Brooks & Wood, as cited by Ahn et al., 2001). Other studies with artificial “natural” kinds have directly compared essentialist and nonessentialist structures but have yielded mixed results (e.g., Rehder & Hastie, 2001). It is possible that explicit training overrides people’s natural tendency to think in terms of a common cause.

In the absence of more direct evidence for essence, the essentialist-minimalist debate is likely to continue (see Ahn et al., 2001; Sloman & Malt, 2003; and Strevens, 2001, for the latest salvos in this dispute). Indeed, the authors of this chapter are not in full agreement. Medin finds minimalism too unconstrained, whereas Rips opines that essentialism suffers from the opposite problem. Adding a predisposition toward parsimony to the minimalist view seems like a constructive move, but such a move would shift minimalism considerably closer to essentialism. Ultimately the issue boils down to determining to what extent causal understandings are biased toward the assumption of a unique, central cause for a category’s usual properties.

SORTALISM

According to some versions of essentialism, an object’s essence determines not only which category it belongs to but also the object’s very identity. According to this view, it is by virtue of knowing that Fido is a dog that you know (in principle) how to identify Fido over time, how to distinguish Fido from other surrounding objects, and how to determine when Fido came into existence and when he will go out of it. In particular, if Fido happens to lose his dog essence, then Fido not only ceases to be a dog, he ceases to exist entirely. As we noted in discussing essentialism, not all categories provide these identity conditions. Being a pet, for example, does not lend identity to Fido, since he may continue to survive in the wild as a non-pet. According to one influential view (Wiggins, 1980), the critical identity-lending category is the one that answers the

question *What is it?* for an object; and since basic-level categories are sometimes defined in just this way, basic-level categories are the presumed source of the principles of identity. (Theories of this type usually assume that identity conditions are associated with just one category for each object, since multiple identity conditions lead to contradictions; see Wiggins, 1980). Contemporary British philosophy tends to refer to such categories as *sortals*, and we will adopt this terminology here.

Sortalism plays an important role in current developmental psychology because developmentalists have used children’s mastery of principles of identity to decide whether these children possess the associated concept. Xu and Carey (1996) staged for infants a scene in which (e.g.) a toy duck appears from one side of an opaque screen and then returns behind it. A toy truck next emerges from the other side of the screen and then returns to its hidden position. Infants habituate after a number of encores of this performance, at which time the screen is removed to reveal both the duck and truck (the scene that adults expect) or just one of the objects (duck or truck). Xu and Carey reported that younger infants (e.g., 10-month-olds) exhibit no more surprise at seeing one object than at seeing two, while older infants (and adults) show more surprise at the one-object tableau. Xu and Carey also showed in control experiments that younger and older infants perform identically if they see a preview of the two starring objects together before the start of the performance. The investigators infer that the younger infants lack the concepts DUCK and TRUCK, since they are unable to use a principle of identity for these concepts to discern that a duck cannot turn into a truck while behind the screen. Xu and Carey’s experiments have sparked a controversy about whether the experimental conditions are simple enough to allow babies to demonstrate their grip on object identity (see Wilcox & Baillargeon, 1998; Xu, 2003), but for present purposes what is important is the assumption that infants’ inability to reidentify objects over temporal gaps implies lack of the relevant concepts.

Sortal theories impose strong constraints on some versions of essentialism. We noted that one of essentialism’s strong points is its ability to explain some of the normative properties of concepts, for example, the role concepts play in inductive inferences. However, sortalism places some restrictions on this ability. Members of sortal categories cannot lose their essence without losing their existence, even in

counterfactual circumstances. This means that if we are faced with a premise like *Suppose dogs can bite through wire...*, we cannot reason about this supposition by assuming that the essence of dogs has changed in such a way as to make dogs stronger. A dog with changed essence is not a superdog, according to sortalism, but rather has ceased to exist (see Rips, 2001). For the same reason, it is impossible to believe without contradiction both that basic-level categories are sortals and that objects can shift from one basic-level category to another.

These consequences of sortalism may be reasonable ones, but it is worth considering the possibility that sortalism—however well it fares as a metaphysical outlook—incorrectly describes people's views about object identity. Although objects typically do not survive a leap from one basic-level category to another, it may not be impossible for them to do so. Blok, Newman, and Rips (2005) and Liitschwager (1995) gave participants scenarios that described novel transformations that sometimes altered the basic-level category. In both studies, participants were more likely to agree that the transformed object was identical to the original if the transformational distance was small. But these judgments could not always be predicted by basic-level membership.

Results from these sci-fi scenarios should be treated cautiously, but they suggest that people think individual objects have an integrity that does not necessarily depend on their basic-level category. Although this idea may be flawed metaphysics, it is not unreasonable as psychometaphysics. People may think that individuals exist as the result of local causal forces, forces that are only loosely tethered to basic-level kinds. As long as these forces continue to support the individual's coherence, that individual can exist even if it finds itself in a new basic-level category. Of course, not all essentialists buy into this link between sortalism and essentialism. For example, people might believe that individuals have both a category essence *and* a set of historical and other characteristics that make it unique. Gutheil and Rosengren (1996) hypothesize that objects have two difference essences, one for membership and another for identity. Just how individual identity and kind identity play out under these scenarios could then be highly variable.

Domain Specificity

The notion of domain specificity has served to organize a great deal of research on conceptual development. For example, much of the work on

essentialism has been conducted in the context of exploring children's naïve biology (see also Au, 1994; Carey, 1995; Gopnik & Wellman, 1994; Spelke, Phillips, & Woodward, 1995). Learning in a given domain may be guided by certain skeletal principles, constraints, and (possibly innate) assumptions about the world (see Gelman & Coley, 1990; Gelman, 2003; Keil, 1981; Kellman & Spelke, 1983; Markman, 1990; Spelke, 1990). Susan Carey's influential (1985) book presented a view of knowledge acquisition as built on framework theories that entail ontological commitments in the service of a causal understanding of real-world phenomena. Two domains can be distinguished from one another if they represent ontologically distinct entities and sets of phenomena and are embedded within different causal explanatory frameworks. These ontological commitments serve to organize knowledge into domains such as naïve physics (or mechanics), naïve psychology, or naïve biology (e.g., see Au, 1994; Carey, 1995; Gelman & Koenig, 2001; Gopnik & Wellman, 1994; Hatano & Inagaki, 1994; Keil, 1994; Spelke et al., 1995; Wellman & Gelman, 1992). In the following we will focus on one candidate domain, naïve biology.

FOLK BIOLOGY AND UNIVERSALS

There is fairly strong evidence that all cultures partition local biodiversity into taxonomies whose basic level is that of the “generic species” (Atran, 1990; Berlin, Breedlove, & Raven, 1973). Generic species often correspond to scientific species (e.g., elm, wolf, and robin); however, for the large majority of perceptually salient organisms (see Hunn, 1999), such as vertebrates and flowering plants, a scientific genus frequently has only one locally occurring species (e.g., bear, oak). In addition to the spontaneous division of local flora and fauna into generic species, cultures seem to structure biological kinds into hierarchically organized groups, such as white oak/oak/tree. Folk-biological ranks vary little across cultures as a function of theories or belief systems (see Malt, 1994, for a review). For example, in studies with Native American and various U.S. and Lowland Maya groups, correlations between folk taxonomies and classical evolutionary taxonomies of the local fauna and flora average $r = .75$ at the generic-species level and about 0.5 with higher levels included (Atran, 1999; Bailenson et al., 2002; Medin et al., 2002). Much of the remaining variance owes to obvious perceptual biases (Itza' Maya group bats with birds in the same life form) and local

ecological concerns. Contrary to received notions about the history and cross-cultural basis for folk-biological classification, utility does *not* appear to drive folk taxonomies (cf. Berlin et al., 1973).

These folk taxonomies also appear to guide and constrain reasoning. For example, Coley, Medin, and Atran (1997) found that both Itza' Maya and U.S. undergraduates privilege the generic-species level in inductive reasoning. That is, an inference from Swamp White Oak to all White Oaks is little if any stronger than an inference from Swamp White Oak to all Oaks. Above the level of Oak, however, inductive confidence takes a sharp drop. In other words, people in both cultures treat the generic level (e.g., Oak) as maximizing induction potential. The results for undergraduates are surprising, since the original Rosch et al. (1976) basic-level studies had suggested that a more abstract level (e.g., TREE) acted as basic for undergraduates and should have been privileged in induction. That is, there is a discrepancy between results with undergraduates on basicness in naming, perceptual classification, and feature listing, on the one hand, and inductive inference, on the other. Coley et al. suggest that the reasoning task relies on expectations rather than knowledge, and that undergraduates may know very little about biological kinds (see also Wolff, Medin, & Pankratz, 1999). Medin and Atran (2004) caution against generalizing results on biological thought from undergraduates, since most have relatively little firsthand experience with nature.

INTERDOMAIN DIFFERENCES

Carey (1985) argued that children initially understand biological concepts like ANIMAL in terms of folk psychology, treating animals as similar to people in having beliefs and desires. Others (e.g., Keil, 1989) argue that young children do have biologically specific theories, albeit more impoverished than those of adults. For example, Springer and Keil (1989) show that preschoolers think biological properties are more likely to be passed from parent to child than are social or psychological properties. They argue that this implies that the children have a biology-like inheritance theory.

The evidence concerning this issue is complex. (see, e.g., Au & Romo, 1996, 1999; Carey, 1991; Coley, 1995; Gelman, 2003; Hatano & Inagaki, 1996; Inagaki, 1997; Inagaki & Hatano, 1993, 1996, 2002; Johnson & Carey, 1998; Keil, 1995; Keil, Levin, Richman, & Gutheil, 1999; Rosengren,

Gelman, Kalish, & McCormick, 1991; Springer & Keil, 1989, 1991). On one hand, Solomon, Johnson, Zaitchik, and Carey (1996) claim that preschoolers do not have a biological concept of inheritance, because they do not have an adult's understanding of the biological causal mechanism involved. On the other hand, there is growing cross-cultural evidence that 4- to 5-year-old children believe (like adults) that the identity of animals and plants follows that of their progenitors, regardless of the environment in which the progeny matures (e.g., progeny of cows raised with pigs, acorns planted with apple trees) (Atran et al., 2001; Gelman & Wellman, 1991; Sousa et al., 2002). Furthermore, it appears that Carey's (1985) results on psychology versus biology may only hold for urban children who have little intimate contact with nature (Atran et al., 2001; Ross et al., 2003). Altogether, the evidence suggests that 4- to 5-year-old children do have a distinct biology, though perhaps one without a detailed model of causal mechanisms (see Rozenblit & Keil, 2002, for evidence that adults also have only a superficial understanding of mechanisms).

DOMAINS AND BRAIN REGIONS

While some cognitive scientists have argued strongly for the idea that different biological categories, such as animals, are represented in different regions of the brain, we saw earlier that the current evidence favors the hypothesis that these categories are represented in terms of their constituent features (the perceptual-functional hypothesis). There is, however, some evidence for different categories being represented in different brain regions, but the categories involved are not the familiar animals and tools, but rather faces and places.

Consider faces. Why should the category of faces get special treatment in the brain? Because faces are special in a number of ways. First, unlike most categories, faces do not seem to be perceived in terms of their constituent features; rather, the visual perception of a face involves configurational processes, not featural ones (e.g., Tanaka & Farah, 1993; Tanaka & Sengco, 1997). Second, the recognition of a face is substantially disrupted by inverting it, more so than with any other category of objects (Yin, 1969). And lastly, recognition of faces may be more critical to an infant's survival than recognition of just about any other class of objects—the first thing that an infant needs to know is who are its conspecifics, as they are the ones who will take care of the infant.

The neural basis of face recognition and face categorization again are located in the general region responsible for object recognition, the ventral temporal cortex. Within this broad region, faces activate a subregion called the fusiform face area, or FFA (so called because the subregion lies on the fusiform gyrus). The precise location of the FFA varies somewhat from person to person, but it is always in the fusiform gyrus of the temporal cortex (Kanwisher, McDermott, & Chun, 1997). Most important, viewing human faces will activate this subregion more than does viewing any other object, including body parts (e.g., Kanwisher, 2000). Indeed, some studies show that the FFA activates to virtually every presentation of a face, but not at all to many other objects, such as letter strings, animal torsos, and the backs of human heads. Furthermore, activation in the FFA is less when viewing inverted than upright faces, which fits perfectly with the fact that inversion disrupts face recognition. That is, if increased activation in the FFA mediates successful face recognition, then unsuccessful face recognition should go with less activation in the FFA.

Some have argued that the special status of faces—the fact that this domain gets its own specialized chunk of brain—is really due to the fact that we spend so much time recognizing faces that we become experts at it, and it is really the expertise that is responsible for the brain specialization (e.g., Gauthier, Tarr, Anderson, Skudlarski, & Gore, 1999). There is something to this point, as researchers have shown that the FFA is increasingly activated by a novel domain of objects as the subject becomes increasingly expert at discriminating among novel objects with a face-like structure (Gauthier, Anderson, Tarr, Skudlarski, & Gore, 1997). Still, neither such highly practiced objects nor even other kinds of faces, like the faces of dogs, will activate the FFA as much as human faces do (Kanwisher, 2000).

There is also evidence that another subregion of ventral temporal cortex is specialized for the domain of places: landscapes and layouts of buildings. The region is referred to as the “parahippocampal place area,” or PPA (as it is located on the parahippocampal gyrus). Viewing scenes of landscapes or buildings will activate the PPA more than will any other class of objects (Epstein, Harris, Stanley, & Kanwisher, 1999). Faces do not activate the PPA just as places do not activate the FFA. And activation of the place area seems sufficient for one to be able to categorize it as a place.

Will other domain-specific areas be found in the brain? There is suggestive evidence of a “word area” located in the occipital cortex of literate people (distinct from the face and place areas) (e.g., Polk & Farah, 2002). What is interesting about this domain-specific area is that it must be acquired through experience with reading. Like the face and place areas, the word area has been studied more in the context of object recognition than object categorization. So at this point in time, there may a limit to what these domain-specific categories can tell us about the implementation of concepts in the human brain.

DOMAINS AND MEMORY

The issue of domain specificity returns us to one of our earlier themes: Does memory organization depend on meaning? We have seen that early research on semantic memory was problematic in this respect, since many of the findings that investigators used to support meaning-based organization had alternative explanations. General-purpose decision processes could produce the same pattern of results, even if the information they operated on was haphazardly organized. Of course, in those olden days, semantic memory was supposed to be a hierarchically organized network like that in Figure 11.1; the network clustered concepts through shared superordinates and properties but was otherwise undifferentiated. Modularity and domain specificity offer a new interpretation of semantic-based memory structure—a partition of memory space into distinct theoretical domains. Can large-scale theories like these support memory organization in a more adequate fashion than homogeneous networks?

One difficulty in merging domain specificity with memory structure is that domain theories do not taxonomize categories; they taxonomize assumptions. What differentiates domains is the set of assumptions or warrants they make available for thinking and reasoning (see Toulmin, 1958, for one such theory), and this means that a particular category of objects usually falls in more than one domain. To put it another way, domain-specific theories are “stances” (Dennett, 1971) or “construals” (Keil, 1995) that overlap in their instances. Take the case of people. The naive psychology domain treats people as having beliefs and goals that lend themselves to predictions about actions (e.g., Leslie, 1987; Wellman, 1990). The naive physics domain treats people as having properties like mass and velocity that warrant predictions about support and

motion (e.g., Clement, 1983; McCloskey, 1983). The naive law-school domain treats people as having properties, such as social rights and responsibilities, that lead to predictions about obedience or deviance (e.g., Fiddick, Cosmides, & Tooby, 2000). The naive biology domain (at least in the Western adult version) treats people as having properties like growth and self-animation that lead to expectations about behavior and development. In short, each ordinary category may belong to many domains.

If domains organize memory, then long-term memory will have to store a concept in each of the domains to which it is related. Such an approach makes some of the difficulties of the old semantic-memory theories yet more perplexing. Recall the issue of identifying the same concept across individuals, which we discussed earlier (see section on “Concepts as Positions in Memory Structure”). Memory modules have the same problem, but they add to it the dilemma of identifying concepts *within* individuals. How do you know that PEOPLE in your psychology module is the same concept as PEOPLE in your physics module and PEOPLE in your law-school module? Similarity is out (since the modules won’t organize them in the same way), spelling is out (both concepts might be tied to the word “people” in an internal dictionary, but then fungi and metal forms are both tied to the word “mold”), and interconnections are out (since they would defeat the idea that memory is organized by domain). We cannot treat the multiple PEOPLE concepts as independent either, since it is important to get back and forth between them. For example, the rights-and-responsibilities information about people in your law-school module has to get together with the goals-and-desires information about people in your psychology module in case you have to decide, together with your fellow jury members, whether the killing was a hate crime or was committed with malice aforethought.

It is reasonable to think that background theories provide premises or grounds for inferences about different topics, and it is also reasonable to think that these theories have their “proprietary concepts.” But if we take domain-specific modules as the basis for memory structure—as a new semantic memory—we have to worry about nonproprietary concepts, too. We have argued that there must be such concepts, since we can reason about the same thing with different theories. Multiple storage is a possibility, if you are willing to forego memory economy and parsimony, and if you can solve the

identifiability problem that we discussed in the previous paragraph. Otherwise, these domain-independent concepts have to inhabit a memory space of their own, and modules cannot be the whole story.

SUMMARY

We seem to be arriving at a skeptical position with respect to the question of whether memory is semantically organized, but we need to be clear about what is and what is not in doubt. What we doubt is that there is compelling evidence that long-term memory is structured in a way that mirrors lexical structure, as in the original semantic-memory models. We do not doubt that memory reflects meaningful relations among concepts, and it is extremely plausible that these relations depend to some extent on word meanings. For example, there may well be a relation in memory that links the concept TRUCKER with the concept BEER, and the existence of this link is probably due in part to the meaning of “trucker” and “beer.” What is not so clear is whether memory structure directly reflects the sort of relations that, in linguistic theory, organizes the meaning of words (where, e.g., “trucker” and “beer” are probably not closely connected). We note, too, that we have not touched (and we do not take sides on) two related issues, which are themselves subjects of controversy.

One of these residual issues is whether there is a split in memory between (a) general knowledge and (b) personally experienced information that is local to time and place. *Semantic memory* (Tulving, 1972) or *generic memory* (Hintzman, 1978) is sometimes used as a synonym for general knowledge in this sense, and it is possible that memory is partitioned along the lines of this semantic/episodic difference, even though the semantic side is not organized by lexical content. The controversy in this case is how such a dual organization can handle learning of “semantic” information from “episodic” encounters (see Tulving, 1984, and his critics in the same issue of *Behavioral and Brain Sciences* for the ins and outs of this debate).

The second issue that we are shirking is whether distributed brands of connectionist models can provide a basis for meaning-based memory. One reason for shirking is that distributed organization means that concepts like DAISY and CUP are *not* stored according to their lexical content. Instead, parts of the content of each concept are smeared across memory in overlapping fashion. It is possible, however, that at a subconcept level—at the level of

features or hidden units—memory has a semantic dimension, and we must leave this question open.

Conclusions and Future Directions

Part of our charge was to make some projections about the future of research on concepts. We do not recommend a solemn attitude toward our predictions. But there are several trends that we have identified and, barring unforeseen circumstances (never a safe assumption), these trends should continue. One property our nominations share is that they uniformly broaden the scope of research on concepts. Here is our shortlist.

Sensitivity to Multiple Functions

The prototypical categorization experiment involves training undergraduates for about an hour and then giving transfer tests to assess what they have learned. This practice is becoming increasingly atypical, even among researchers studying artificially constructed categories in the lab. Recently researchers have studied functions other than categorization, as well as interactions across functions (see also Solomon et al., 1999).

Broader Applications of Empirical Generalizations and Computational Models

As a wider range of conceptual functions come under scrutiny, new generalizations emerge and computational models face new challenges (e.g., Yamauchi et al., 2002). Both developments set the stage for better bridging to other contexts and applications. This is perhaps most evident in the area of cognitive neuroscience where computational models have enriched studies of multiple categorization and memory systems (and vice versa). Norman, Brooks, Coblenz, and Babcock (1992) provide a nice example of extensions from laboratory studies to medical diagnosis in the domain of dermatology.

Greater Interactions Between Work on Concepts and Psycholinguistic Research

We have pressed the point that research on concepts has diverged from psycholinguistics because two different concepts of concepts seem to be in play in these fields. But it cannot be true that the concepts we use in online sentence understanding are unrelated to the concepts we employ in reasoning and categorizing. There is an opportunity for theorists and experimenters here to provide an account of the interface between these functions. To our knowledge, no one has tried to build a model

that would integrate concept combination over its different stages, but it might be a worthwhile goal to do so. One possibility, for example, is to use sentence comprehension techniques to track the way that the lexical content of a word in speech or text is transformed in deeper processing (see Piñango, Zurif, & Jackendoff, 1999, for one effort in this direction). Another type of effort at integration is Wolff and Song's (2003) work on causal verbs and people's perception of cause, which contrasts predictions derived from cognitive linguistics with those from cognitive psychology.

Greater Diversity of Participant Populations

Although research with undergraduates at major Western universities will probably never go out of style (precedent and convenience are two powerful staying forces), we expect the recent increase to continue in the use of other populations. Work by Nisbett and his associates (e.g., Nisbett, Peng, Choi, & Norenzayan, 2001; Nisbett & Norenzayan, 2002) has called into question the idea that basic cognitive processes are universal, and categories and conceptual functions are basic cognitive functions. In much of the work by Atran, Medin, and their associates, undergraduates are the “odd group out” in the sense that their results deviate from those of other groups. In addition, cross-linguistic studies are often an effective research tool for addressing questions about the relationship between linguistic and conceptual development (e.g., Waxman, 1999).

More Psychometaphysics

An early critique of the “theory theory” is that it suffered from vagueness and imprecision. As we have seen in this review, however, this framework has led to more specific claims (e.g., Ahn’s causal status hypothesis) and the positions are clear enough to generate theoretical controversies (e.g., contrast Smith, Jones, & Landau, 1996, with Gelman, 2000; Booth & Waxman, 2002, 2003, with Smith, Jones, Yoshida, & Colunga, 2003; and Ahn, 1998, with Rehder & Kim, 2006 and Rehder, 2009). It is safe to predict even greater future interest in these questions.

All of the Above in Combination

Concepts and categories are shared by all the cognitive sciences, so there is very little room for researchers to stake out a single paradigm or sub-topic and work in blissful isolation. Although the

idea of a semantic memory uniting memory structure, lexical organization, and categorization may have been illusory, this does not mean that progress will be fostered by ignoring the insights on concepts that these perspectives (and others) provide. We may see further fragmentation in the concepts of concepts, but it will still be necessary to explore the relations among them. Our only firm prediction is that the work we will find most exciting will be research that draws on multiple points of view.

References

- Ahn, W.-K. (1998). Why are different features central for natural kinds and artifacts? The role of causal status in determining feature centrality. *Cognition*, 69, 135–178.
- Ahn, W.-K., Kalsish, C., Gelman, S. A., Medin, D. L., Luhmann, C., Atran, S., & Shafto, P. (2001). Why essences are essential in the psychology of concepts. *Cognition*, 82, 59–69.
- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.
- Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, 14, 471–517.
- Anderson, J. R., & Bower, G. H. (1973). *Human associative memory*. Hillsdale, NJ: Erlbaum.
- Anderson, J. R., & Fincham, J. M. (1996). Categorization and sensitivity to correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 259–277.
- Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, 37, 372–400.
- Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, 105, 442–481.
- Atran, S. (1990). *Cognitive foundations of natural history*. Cambridge, UK: Cambridge University Press.
- Atran, S. (1998). Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Sciences*, 21, 547–609.
- Atran, S. (1999). Itzaj Maya folk-biological taxonomy. In D. Medin & S. Atran (Eds.), *Folk biology* (pp. 119–204). Cambridge, MA: MIT Press.
- Atran, S., Medin, D., Lynch, E., Vapnarsky, V., Ucan Ek', E., & Sousa, P. (2001). Folkbiology doesn't come from folkpsychology: Evidence from Yukatek Maya in cross-cultural perspective. *Journal of Cognition and Culture*, 1, 3–42.
- Au, T. K.-f. (1994). Developing an intuitive understanding of substance kinds. *Cognitive Psychology*, 27, 71–111.
- Au, T. K.-f., & Romo, L. F. (1996). Building a coherent conception of HIV transmission: A new approach to AIDS education. In D. L. Medin (Ed.), *The psychology of learning and motivation: Advances in research and theory*, Vol. 35. (pp. 193–241). San Diego, CA: Academic Press.
- Au, T. K.-f., & Romo, L. F. (1999). Mechanical causality in children's 'Folkbiology'. In D. L. Medin & S. Atran (Eds.) *Folkbiology*. (pp. 355–401). Cambridge, MA: The MIT Press.
- Bailenson, J. N., Shum, M., Atran, S., Medin, D. L., & Coley, J. D. (2002). A bird's eye view: Biological categorization and reasoning within and across cultures. *Cognition*, 84, 1–53.
- Balota, D. A. (1994). Visual word recognition: A journey from features to meaning. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 303–358). San Diego, CA: Academic Press.
- Barsalou, L. W. (1983). Ad-hoc categories. *Memory and Cognition*, 11, 211–227.
- Berlin, B., Breedlove, D., & Raven, P. (1973). General principles of classification and nomenclature in folk biology. *American Anthropologist*, 75, 214–242.
- Blok, S., Newman, G., & Rips, L. J. (2005). Individuals and their concepts. In W.-k. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), *Categorization inside and outside the lab* (pp. 127–149). Washington, D.C.: American Psychological Association.
- Bloom, P. (1996). Intention, history, and artifact concepts. *Cognition*, 60, 1–29.
- Booth, A. E., & Waxman, S. R. (2002). Object names and object functions serve as cues to categories for infants. *Developmental Psychology*, 38, 948–957.
- Booth, A. E., & Waxman, S. R. (2003). Bringing theories of word learning in line with the evidence. *Cognition*, 87, 215–218.
- Bourne, L. E., Jr. (1970). Knowing and using concepts. *Psychological Review*, 77, 546–556.
- Bozoki, A., Grossman, M., & Smith, E. E. (2006). Can patients with Alzheimer's disease learn a category implicitly? *Neuropsychologia*, 44, 816–827.
- Braiby, N., Franks, B., & Hampton, J. (1996). Essentialism, word use, and concepts. *Cognition*, 59, 247–274.
- Brooks, L. R. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and Categorization* (pp. 169–211). New York: Wiley.
- Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.
- Buckner, R. L. (2000). Neuroimaging of memory. In M. Gazzaniga (Ed.), *The new cognitive neurosciences* (2nd eEd.) (pp. 1013–1022). Cambridge, MA: MIT Press.
- Burge, T. (1999). Comprehension and interpretation. In L. E. Hahn (Ed.), *The philosophy of Donald Davidson* (pp. 229–250). Chicago, IL: Open Court.
- Busemeyer, J. R., Dewey, G. I., & Medin, D. L. (1984). Evaluation of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10, 638–648.
- Caramazza, A., & Grober, E. (1976). Polysemy and the structure of the subjective lexicon. In C. Rameh (Ed.), *Georgetown University round table on language and linguistics* (pp. 181–206). Washington, D.C.: Georgetown University Press.
- Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.
- Carey, S. (1991). Knowledge acquisition: Enrichment or conceptual change? In S. Carey & R. Gelman (Eds.), *The epigenesis of mind: Essays on biology and cognition* (pp. 257–291). Hillsdale, NJ, England: Lawrence Erlbaum Associates Inc.
- Carey, S. (1995). On the origin of causal understanding. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 268–308). New York: Oxford University Press.
- Chao, L. L., & Martin, A. (1999). Cortical regions associated with perceiving, naming, and knowing about colors. *Journal of Cognitive Neuroscience*, 11, 25–35.

- Chierchia, G., & McConnell-Ginet, S. (1990). *Meaning and grammar: An introduction to semantics*. Cambridge, MA: MIT Press.
- Clapper, J., & Bower, G. (2002). Adaptive categorization in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 908–923.
- Clement, J. (1983). A conceptual model discussed by Galileo and used intuitively by physics students. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 325–340). Hillsdale, NJ: Erlbaum.
- Coley, J. D. (1995). Emerging differentiation of folkbiology and folkpsychology: Attributions of biological and psychological properties to living things. *Child Development*, 66, 1856–1874.
- Coley, J. D., Medin, D. L., & Atran, S. (1997). Does rank have its privilege? Inductive inferences within folkbiological taxonomies. *Cognition*, 64, 73–112.
- Collins, A. M., & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review*, 82, 407–428.
- Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8, 240–247.
- Conrad, F. G., & Rips, L. J. (1986). Conceptual combination and the given/new distinction. *Journal of Memory and Language*, 25, 255–278.
- Dennett, D. C. (1971). Intensional systems. *Journal of Philosophy*, 68, 87–106.
- Diesendruck, G., & Gelman, S. A. (1999). Domain differences in absolute judgments of category membership: Evidence for an essentialist account of categorization. *Psychonomic Bulletin & Review*, 6, 338–346.
- Epstein, R., Harris, A., Stanley, D., & Kanwisher, N. (1999). The parahippocampal place area: Recognition, navigation, or encoding? *Neuron*, 23, 115–125.
- Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, 127, 107–140.
- Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology*, 18, 500–549.
- Farah, M. J., & McClelland, J. L. (1991). A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of Experimental Psychology: General*, 120, 339–357.
- Fiddick, L., Cosmides, L., & Tooby, J. (2000). No interpretation without representation: The role of domain-specific representations and inferences in the Wason selection task. *Cognition*, 77, 1–79.
- Fillmore, C. J., & Atkins, B. T. S. (2000). Describing polysemy: The case of ‘crawl.’ In Y. Ravin & C. Leacock (Eds.), *Polysemy: Theoretical and computational approaches* (pp. 91–110). Oxford, England: Oxford University Press.
- Filoteo, J. V., Maddox, W. T., & Davis, J. D. (2001). A possible role of the striatum in linear and nonlinear categorization rule learning: Evidence from patients with Huntington’s disease. *Behavioral Neuroscience*, 115, 786–798.
- Fodor, J. (1994). Concepts: A portboiler. *Cognition*, 50, 95–113.
- Fodor, J. (1998). *Concepts: Where cognitive science went wrong*. Oxford, England: Oxford University Press.
- Franks, B. (1995). Sense generation: A “quasi-classical” approach to concepts and concept combination. *Cognitive Science*, 19, 441–505.
- Furnham, A. (1995). Lay beliefs about phobia. *Journal of Clinical Psychology*, 51, 518–525.
- Gagné, C. L., & Shoben, E. J. (1997). Influence of thematic relations on the comprehension of modifier-head combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 71–87.
- Gauthier, I., Anderson, A. W., Tarr, M. J., Skudlarski, P., & Gore, J. C. (1997). Levels of categorization in visual object recognition studied with functional MRI. *Current Biology*, 7, 645–651.
- Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform ‘face area’ increases with expertise in recognizing novel objects. *Nature*, 2, 568–573.
- Gelman, S. A. (2000). The role of essentialism in children’s concepts. In H. W. Reese (Ed.), *Advances in child development and behavior* (Vol. 27, pp. 55–98). San Diego, CA: Academic Press.
- Gelman, S. A. (2003). *The essential child: origins of essentialism in everyday thought*. New York: Oxford University Press.
- Gelman, S. A., & Coley, J. D. (1990). The importance of knowing a dodo is a bird: Categories and inferences in 2-year-old children. *Developmental Psychology*, 26(5), 796–804.
- Gelman, S. A., & Hirschfeld, L. A. (1999). How biological is essentialism? In D. L. Medin & S. Atran (Eds.), *Folkbiology* (pp. 403–446). Cambridge, MA: MIT Press.
- Gelman, S. A., & Koenig, M. A. (2001). The role of animacy in children’s understanding of “move.” *Journal of Child Language*, 28, 683–701.
- Gelman, S. A., & Wellman, H. M. (1991). Insides and essence: Early understandings of the non-obvious. *Cognition*, 38, 213–244.
- Gelman, S. A., Star, J. R., & Flukes, J. E. (2002). Children’s use of generics in inductive inference. *Journal of Cognition and Development*, 3, 179–199.
- Ghiselin, M. T. (1981). Categories, life, and thinking. *Behavioral and Brain Sciences*, 4, 269–313.
- Gleitman, L. R., & Gleitman, H. (1970). *Phrase and paraphrase*. New York: W.W. Norton.
- Goldstone, R. L. (2003). Learning to perceive while perceiving to learn. In R. Kimchi, M. Behrmann, & C. Olson (Eds.), *Perceptual organization in vision: Behavioral and neural perspectives* (pp. 233–278). Mahwah, NJ: Erlbaum.
- Goldstone, R. L., & Rogosky, B. J. (2002). Using relations within conceptual systems to translate across conceptual systems. *Cognition*, 84, 295–320.
- Goldstone, R. L., & Stevvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General*, 130, 116–139.
- Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, 78, 27–43.
- Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, 49, 585–612.
- Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 257–293). New York, NY: Cambridge University Press.
- Gutheil, G., & Rosengren, K. S. (1996). A rose by any other name: Preschoolers’ understanding of individual identity across name and appearance changes. *British Journal of Developmental Psychology*, 14, 477–498.
- Hampton, J. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 18, 441–461.

- Hampton, J. (1987). Inheritance of attributes in natural concept conjunctions. *Memory and Cognition*, 15, 55–71.
- Hampton, J. (1997). Conceptual combination. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts, and categories* (pp. 133–159). Cambridge, MA: MIT Press.
- Hastie, R., Schroeder, C., & Weber, R. (1990). Creating complex social conjunction categories from simple categories. *Bulletin of the Psychonomic Society*, 28, 242–247.
- Hatano, G., & Inagaki, K. (1994). Young children's naive theory of biology. *Cognition*, 50, 171–188.
- Hatano, G., & Inagaki, K. (1996). Cognitive and cultural factors in the acquisition of intuitive biology. In D. R. Olson & N. Torrance (Eds.), *The handbook of education and human development: New models of learning, teaching and schooling* (pp. 683–708). Malden: Blackwell Publishing.
- Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin and Review*, 7, 569–592.
- Hintzman, D. L. (1978). *The psychology of learning and memory*. San Francisco, CA: Freeman.
- Hintzman, D. L. (1986). 'Schema abstraction' in a multiple-trace memory model. *Psychological Review*, 93, 411–428.
- Hollander, M. A., Gelman, S. A., & Star, J. (2002). Children's interpretation of generic noun phrases. *Developmental Psychology*, 38, 883–894.
- Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory*, 7, 418–439.
- Hull, D. (1999). Interdisciplinary dissonances. In D. L. Medin & S. Atran (Eds.), *Folkbiology* (pp. 477–500). Cambridge, MA: MIT Press.
- Hunn, E. (1999). Size as limiting the recognition of biodiversity in folkbiological classification: One of four factors governing the cultural recognition of biological taxa. In D. L. Medin & S. Atran (Eds.), *Folkbiology* (pp. 47–69). Cambridge, MA: MIT Press.
- Inagaki, K. (1997). Emerging distinctions between naive biology and naive psychology. In H. M. Wellman & K. Inagaki (Eds.), *The emergence of core domains of thought: Children's reasoning about physical, psychological, and biological phenomena* (pp. 27–44). San Francisco, CA: Jossey-Bass.
- Inagaki, K., & Hatano, G. (1993). Young children's understanding of the mind-body distinction. *Child Development*, 64, 1534–1549.
- Inagaki, K., & Hatano, G. (1996). Young children's recognition of commonalities between animals and plants. *Child Development*, 67, 2823–2840.
- Inagaki, K., & Hatano, G. (2002). *Young Children's Naive Thinking about the Biological World*. New York: Psychology Press.
- Johansen, M. J., & Palmeri, T. J. (2002). Are there representational shifts in category learning? *Cognitive Psychology*, 45, 482–553.
- Johnson, C., & Keil, F. (2000). Explanatory understanding and conceptual combination. In F. C. Keil & R. A. Wilson (Eds.), *Explanation and cognition* (pp. 327–359). Cambridge, MA: MIT Press.
- Johnson, S. C., & Carey, S. (1998). Knowledge enrichment and conceptual change in folkbiology: Evidence from Williams syndrome. *Cognitive Psychology*, 37, 156–200.
- Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): A 'lazy' algorithm for probabilistic inference from generic knowledge. *Cognitive Science: A Multidisciplinary Journal*, 26, 563–607.
- Kalish, C. W. (1995). Essentialism and graded membership in animal and artifact categories. *Memory & Cognition*, 23, 335–353.
- Kalish, C. W. (2002). Essentialist to some degree: Beliefs about the structure of natural kind categories. *Memory & Cognition*, 30, 340–352.
- Kan, I. P., Barsalou, L. W., Solomon, K.O., Minor, J. K., & Thompson-Schill, S. L. (2003). Role of mental imagery in a property verification task: fMRI evidence for perceptual representations of conceptual knowledge. *Cognitive Neuropsychology*, 20, 525–540.
- Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience*, 3, 759–763.
- Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17, 4302–4311.
- Keil, F. C. (1981). Constraints on knowledge and cognitive development. *Psychological Review*, 88, 197–227.
- Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.
- Keil, F. C. (1994). The birth and nurturance of concepts by domains: The origins of concepts of living things. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 234–254). Cambridge, UK: Cambridge University Press.
- Keil, F. C. (1995). The growth of causal understanding of natural kinds. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition* (pp. 234–262). Oxford, England: Oxford University Press.
- Keil, F. C., Levin, D. T., Richman, B. A., & Gutheil, G. (1999). Mechanism and explanation in the development of biological thought: The case of disease. In D. Medin & S. Atran (Eds.), *Folkbiology* (pp. 285–319). Cambridge, MA: MIT Press.
- Kellenbach, M. L., Brett, M., & Patterson, K. (2001). Large, colorful, or noisy? Attribute- and modality-specific activations during retrieval of perceptual attribute knowledge. *Cognitive, Affective and Behavioral Neuroscience*, 1, 207–221.
- Kellman, P. J., & Spelke, E. S. (1983). Perception of partly occluded objects in infancy. *Cognitive Psychology*, 15, 483–524.
- Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, 116, 20–58.
- Kim, N. S., & Ahn, W.-k. (2002a). Clinical psychologists' theory-based representations of mental disorders predict their diagnostic reasoning and memory. *Journal of Experimental Psychology: General*, 131, 451–476.
- Kim, N. S., & Ahn, W.-k. (2002b). The influence of naive causal theories on lay concepts of mental illness. *The American Journal of Psychology*, 115, 33–65.
- Klein, D. E., & Murphy, G. L. (2002). Paper has been my ruin: Conceptual relations of polysemous senses. *Journal of Memory and Language*, 47, 548–570.
- Knapp, A. G., & Anderson, J. A. (1984). Theory of categorization based on distributed memory storage. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 10, 616–637.
- Knowlton, B. J., & Squire, L. R. (1993). The learning of categories: Parallel brain systems for item memory and category knowledge. *Science*, 262, 1747–1749.
- Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science*, 273, 1399–1402.

- Krifka, M., Peltier, F. J., Carlson, G. N., ter Meulen, A., Link, G., & Chierchia, G. (1995). Genericity: An introduction. In G. N. Carlson & F. J. Pelletier (Eds.), *The generic book* (pp. 1–124). Chicago, IL: University of Chicago Press.
- Kruschke, J. K. (1992). ALCOVE: An exemplar based connectionist model of category learning. *Psychological Review*, 99, 22–44.
- Kunda, Z., Miller, D. T., & Claire, T. (1990). Combining social concepts: The role of causal reasoning. *Cognitive Science*, 14, 551–577.
- Lamberts, K. (1995). Categorization under time pressure. *Journal of Experimental Psychology: General*, 124, 161–180.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211–240.
- Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind." *Psychological Review*, 94, 412–426.
- Leslie, S.-J. (2008). Generics: Cognition and acquisition. *Philosophical Review*, 117, 1–47.
- Levine, M. (1971). Hypothesis theory and nonlearning despite ideal SR-Reinforcement contingencies. *Psychological Review*, 78, 130–140.
- Liitschwager, J. C. (1995). Children's reasoning about identity across transformations. *Dissertation Abstracts International*, 55 (10), 4623B. (UMI No. 9508399).
- Love, B. C., Markman, A. B., & Yamauchi, T. (2000). Modeling classification and inferencelearning. *Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, 17, 136–141.
- Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review*, 111, 309–332.
- Lucas, M. (2000). Semantic priming without association. *Psychonomic Bulletin and Review*, 7, 618–630.
- Lyons, J. (1977). *Semantics* (Vol. 2). Cambridge, England: Cambridge University Press.
- Maddox, W. T. (1999). On the dangers of averaging across observers when comparing decision bound models and generalized context models of categorization. *Perception and Psychophysics*, 61, 354–375.
- Maddox, W. T. (2002). Learning and attention in multidimensional identification, and categorization: Separating low-level perceptual processes and high level decisional processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 99–115.
- Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception and Psychophysics*, 53, 49–70.
- Maddox, W. T., & Ashby, F. G. (1998). Selective attention and the formation of linear decision boundaries: Comment on McKinley and Nosofsky (1996). *Journal of Experimental Psychology: Human Perception and Performance*, 24, 301–321.
- Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology: Paris*, 102, 59–70.
- Malt, B. C. (1994). Water is not H₂O. *Cognitive Psychology*, 27, 41–70.
- Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior*, 23, 250–269.
- Malt, B. C., Ross, B. H., & Murphy, G. L. (1995). Predicting features for members of natural categories when categorization is uncertain. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 21, 646–661.
- Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing vs. naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, 40, 230–262.
- Markman, A. B., & Makin, V. S. (1998). Referential communication and category acquisition. *Journal of Experimental Psychology: General*, 127, 331–354.
- Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science*, 14, 57–77.
- Martin, A., Haxby, J. V., Lalonde, F. M., Wiggs, C. L., & Ungerleider, L. G. (1995). Discrete cortical regions associated with knowledge of color and knowledge of action. *Science*, 270, 102–105.
- McCloskey, M. (1983). Naive theories of motion. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 299–324). Hillsdale, NJ: Erlbaum.
- McCloskey, M., & Glucksberg, S. (1979). Decision processes in verifying category membership statements: Implications for models of semantic memory. *Cognitive Psychology*, 11, 1–37.
- McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill defined category structures. *Journal of Experimental Psychology: Human Perception and Performance*, 21, 128–148.
- Medin, D. (1989). Concepts and conceptual structures. *American Psychologist*, 45, 1469–1481.
- Medin, D. L., & Atran, S. (2004). The native mind: Biological categorization and reasoning in development and across cultures. *Psychological Review*, 111, 960–983.
- Medin, D. L., & Coley, J. D. (1998). Concepts and categorization. In J. Hochberg (Ed.), *Handbook of perception and cognition. Perception and cognition at century's end: History, philosophy, theory* (pp. 403–439). San Diego, CA: Academic Press.
- Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou and A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–195). New York: Cambridge University Press.
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85, 207–238.
- Medin, D. L., & Shoben, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology*, 20, 158–190.
- Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8, 37–50.
- Medin, D. L., Lynch, E. B., & Solomon, K. O. (2000). Are there kinds of concepts? *Annual Review of Psychology*, 51, 121–147.
- Medin, D. L., Ross, N., Atran, S., Burnett, R. C., & Blok, S. V. (2002). Categorization and reasoning in relation to culture and expertise. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 41, pp. 1–41). Amsterdam: Elsevier.
- Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90, 227–234.
- Miller, G. A., & Johnson-Laird, P. N. (1976). Language and perception. Cambridge, MA: Harvard University Press.

- Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 775–799.
- Murphy, G. L. (1988). Comprehending complex concepts. *Cognitive Science*, 12, 529–562.
- Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.
- Murphy, G. L., & Ross, B. H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology*, 27, 148–193.
- Nisbett, R. E., & Norenzayan, A. (2002). Culture and cognition. In H. Pashler, H. & D. Medin, D. (Eds.), *Strevens' handbook of experimental psychology. Vol. 2: Memory and cognitive processes* (3rd ed., pp. 561–597). New York: Wiley.
- Nisbett, R., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic vs. analytic cognition. *Psychological Review*, 108, 291–310.
- Norman, D. A., & Rumelhart, D. E. (1975). *Explorations in cognition*. San Francisco, CA: W. H. Freeman.
- Norman, G. R., Brooks, L. R., Coblenz, C. L., & Babcock, C. J. (1992). The correlation of feature identification and category judgments in diagnostic radiology. *Memory & Cognition*, 20, 344–355.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115, 39–57.
- Nosofsky, R. M. (1991). Tests of an exemplar model for relation perceptual classification and recognition in memory. *Journal of Experimental Psychology: Human Perception and Performance*, 17, 3–27.
- Nosofsky, R. M. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar-based interpretation. *Psychological Science*, 9, 247–255.
- Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar based accounts of “multiple system” phenomena in perceptual categorization. *Psychonomic Bulletin and Review*, 7, 375–402.
- Nosofsky, R. M., & Palmeri, T. J. (1997a). An exemplar based random walk model of speeded classification. *Psychological Review*, 104, 266–300.
- Nosofsky, R. M., & Palmeri, T. J. (1997b). Comparing exemplar retrieval and decision-bound models of speeded perceptual classification. *Perception and Psychophysics*, 59, 1027–1048.
- Nosofsky, R. M., & Zaki, S. R. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar based interpretation. *Psychological Science*, 9, 247–255.
- Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 924–940.
- Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 282–304.
- Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, 101, 53–79.
- Osherson, D. N., & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, 11, 35–58.
- Osherson, D. N., & Smith, E. E. (1982). Gradedness and conceptual combination. *Cognition*, 12, 299–318.
- Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 23, 324–354.
- Palmeri, T. J. (1999). Learning hierarchically structured categories: A comparison of category learning models. *Psychonomic Bulletin and Review*, 6, 495–503.
- Palmeri, T. J., & Flanery, M. A. (1999). Learning about categories in the absence of training: Profound amnesia and the relationship between perceptual categorization and recognition memory. *Psychological Science*, 10, 526–530.
- Palmeri, T. J., & Flanery, M. A. (2002). Memory systems and perceptual categorization. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 41, pp. 141–190). Amsterdam: Elsevier.
- Partee, B. H. (1995). Lexical semantics and compositionality. In D. N. Osherson (Series Ed.) & L. R. Gleitman & M. Liberman (vol. eds.) & D. N. Osherson (series ed.), *Invitation to cognitive science* (vol. 1: *Language*, pp. 311–360). Cambridge, MA: MIT Press.
- Piñango, M. M., Zurif, E., & Jackendoff, R. (1999). Real-time processing implications of enriched composition at the syntax-semantics interface. *Journal of Psycholinguistics Research*, 28, 395–414.
- Polk, T. A., & Farah, M. J. (2002). Functional MRI evidence for an abstract, not perceptual, word-form area. *Journal of Experimental Psychology: General*, 131, 65–72.
- Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, 77, 353–363.
- Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology*, 83, 304–308.
- Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Sciences*, 12, 410–430.
- Quillian, M. R. (1969). The teachable language comprehender: A simulation program and theory of language. *Communications of the ACM*, 12, 459–476.
- Reagher, G., & Brooks, L. R. (1993). Perceptual manifestations of an analytic structure: The priority of holistic individuation. *Journal of Experimental Psychology: General*, 122, 92–114.
- Reber, P. J., Gitelman, D. R., Parrish, T. B., & Mesulam, M. M. (2003). Dissociating explicit and implicit category knowledge with fMRI. *Journal of Cognitive Neuroscience*, 15, 574–583.
- Reber, P. J., Stark, C. E. L., & Squire, L. R. (1998a). Cortical areas supporting category learning identified using functional MRI. *Proceedings of the National Academy of Sciences of the USA*, 95, 747–750.
- Reber, P. J., Stark, C. E. L., & Squire, L. R. (1998b). Contrasting cortical activity associated with category memory and recognition memory. *Learning and Memory*, 5, 420–428.
- Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, 3, 382–407.
- Reed, J. M., Squire, L. R., Patalano, A. L., Smith, E. E., & Jonides, J. J. (1999). Learning about categories that are defined by object-like stimuli despite impaired declarative memory. *Behavioral Neuroscience*, 113, 411–419.
- Rehder, B. (2009). Causal-based property generalization. *Cognitive Science*, 33, 301–343.
- Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General*, 130, 323–360.
- Rehder, B., & Kim, S. (2006). How causal knowledge affects classification: A generative theory of categorization. *Journal*

- of Experimental Psychology: Learning, Memory, and Cognition*, 32, 659–683.
- Restle, F. (1962). The selection of strategies in cue learning. *Psychological Review*, 69, 329–343.
- Rips, L. J. (1995). The current status of research on concept combination. *Mind and Language*, 10, 72–104.
- Rips, L. J. (2001). Necessity and natural categories. *Psychological Bulletin*, 127, 827–852.
- Rips, L. J. (2011). *Lines of thought: Central concepts in cognitive psychology*. Oxford, England: Oxford University Press.
- Rips, L. J., Smith, E. E., & Shoben, E. J. (1978). Semantic composition in sentence verification. *Journal of Verbal Learning & Verbal Behavior*, 17, 375–401.
- Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.
- Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (pp. 111–144). New York: Academic Press.
- Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104, 192–233.
- Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Erlbaum.
- Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7, 573–605.
- Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8, 382–439.
- Rosengren, K. S., Gelman, S. A., Kalish, C. W., & McCormick, M. (1991). As time goes by: Children's early understanding of growth in animals. *Child Development*, 62, 1302–1320.
- Ross, B. H. (1997). The use of categories affects classification. *Journal of Memory and Language*, 37, 240–267.
- Ross, B. H. (1999). Postclassification category use: The effects of learning to use categories after learning to classify. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 25, 743–757.
- Ross, B. H. (2000). The effects of category use on learned categories. *Memory and Cognition*, 28, 51–63.
- Ross, B. H., & Murphy, G. L. (1996). Category based predictions: Influence of uncertainty and feature associations. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22, 736–753.
- Ross, N., Medin, D., Coley, J. D., & Atran, S. (2003). Cultural and experimental differences in the development of folkbiological induction. *Cognitive Development*, 18, 25–47.
- Rozentblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, 26, 521–562.
- Schyns, P., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 23, 681–696.
- Schyns, P., Goldstone, R., & Thibaut, J. (1998). Development of features in object concepts. *Behavioral and Brain Sciences*, 21, 1–54.
- Sloman, S. A., & Malt, B. (2003). Artifacts are not ascribed essences, nor are they treated as belonging to kinds. *Language and Cognitive Processes*, 18, 563–582.
- Sloman, S. A., Love, B. C., & Ahn, W.-K. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, 22, 189–228.
- Smith, E. E. (2008). The case for implicit category learning. *Cognitive, Affective, and Behavioral Neuroscience*, 8, 3–16.
- Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.
- Smith, E. E., Osherson, D. N., Rips, L. J., & Keane, M. (1988). Combining prototypes: A selective modification model. *Cognitive Science*, 12, 485–527.
- Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, 81, 214–241.
- Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 24, 1411–1436.
- Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 26, 3–27.
- Smith, J. D., Murray, M. J., Jr., & Minda, J. P. (1997). Straight talk about linear separability. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 23, 659–680.
- Smith, L. B., Jones, S. S., Yoshida, H., & Colunga, E. (2003). Whose DAM account? Attentional learning explains Booth and Waxman. *Cognition*, 87, 209–213.
- Smith, L. B., Jones, S. S., & Landau, B. (1996). Naming in young children: A dumb attentional mechanism? *Cognition*, 60, 143–171.
- Sober, E. (1980). Evolution, population thinking, and essentialism. *Philosophy of Science*, 47, 350–383.
- Solomon, G. E. A., Johnson, S. C., Zaitchik, D., & Carey, S. (1996). Like father, like son: Young children's understanding of how and why offspring resemble their parents. *Child Development*, 67, 151–171.
- Solomon, K. O., Medin, D. L., & Lynch, E. B. (1999). Concepts do more than categorize. *Trends in Cognitive Science*, 3, 99–105.
- Sousa, P., Atran, S., & Medin, D. (2002). Essentialism and folkbiology: Evidence from Brazil. *Journal of Cognition and Culture*, 2, 195–223.
- Spelke, E. S. (1990). Principles of object perception. *Cognitive Science*, 14, 29–56.
- Spelke, E. S., Phillips, A., & Woodward, A. L. (1995). Infants' knowledge of object motion and human action. In D. Sperber, D. Premack & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 44–78). New York, NY: Clarendon Press/Oxford University Press.
- Springer, K., & Keil, F. C. (1989). On the development of biologically specific beliefs: The case of inheritance. *Child Development*, 60, 637–648.
- Springer, K., & Keil, F. C. (1991). Early differentiation of causal mechanisms appropriate to biological and nonbiological kinds. *Child Development*, 62, 767–781.
- Squire, L. R., & Knowlton, B. J. (1995). Memory, hippocampus, and brain systems. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 825–837). Cambridge, MA: The MIT Press.
- Stanton, R., Nosofsky, R. M., & Zaki, S. (2002). Comparisons between exemplar similarity and mixed prototype models using a linearly separable category structure. *Memory and Cognition*, 30, 934–944.
- Storms, G., de Boeck, P., van Mechelen, I., & Ruts, W. (1998). No guppies, nor goldfish, but tumble dryers, Noriega, Jesse

- Jackson, panties, car crashes, bird books, and Stevie Wonder. *Memory and Cognition*, 26, 143–145.
- Strevens, M. (2000). The essentialist aspect of naïve theories. *Cognition*, 74, 149–175.
- Strevens, M. (2001). Only causation matters. *Cognition*, 82, 71–76.
- Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology*, 46 A, 225–245.
- Tanaka, J. W., & Sengco, J. A. (1997). Features and their configuration in face recognition. *Memory and Cognition*, 25, 583–589.
- Thompson-Schill, S. L. (2003). Neuroimaging studies of semantic memory: Inferring “how” from “where”. *Neuropsychologia*, 41, 280–292.
- Thompson-Schill, S. L., D’Esposito, M., Aguirre, G. K., & Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: A reevaluation. *Proceedings of the National Academy of Sciences USA*, 94, 14792–14797.
- Toulmin, S. (1958). *The uses of argument*. Cambridge, England: Cambridge University Press.
- Trabasso, T., & Bower, G. H. (1968). *Attention in learning: theory and research*. New York: Wiley.
- Tulving, E. (1972). Episodic and semantic memory. In E. Tulving and W. Donaldson (Eds.), *Organization of memory* (pp. 381–403). New York: Academic Press.
- Tulving, E. (1984). Precis of elements of episodic memory. *Behavioral and Brain Sciences*, 7, 223–268.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327–352.
- Verguts, T., Storms, G., & Tuerlinckx, F. (2003). Decision bound theory and the influence of familiarity. *Psychonomic Bulletin and Review*, 10, 141–148.
- Wagner, A. D., Bunge, S. A., & Badre, D. (2004). Cognitive control, semantic memory, and priming: Contributions from prefrontal cortex. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (3rd ed., pp. 709–725). Cambridge, MA: MIT Press.
- Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain*, 107, 829–854.
- Waxman, S. R. (1999). The dubbing ceremony revisited: Object naming and categorization in infancy and early childhood. In D. L. Medin & S. Atran (Eds.), *Folkbiology* (pp. 233–284). Cambridge, MA: MIT Press.
- Wellman, H. M. (1990). *The child’s theory of mind*. Cambridge, MA: MIT Press.
- Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, 43, 337–375.
- Wiggins, D. (1980). *Sameness and substance*. Cambridge, MA: Harvard University Press.
- Wilcox, T., & Baillargeon, R. (1998). Object individuation in infancy: The use of featural information in reasoning about occlusion events. *Cognitive Psychology*, 37, 97–155.
- Wisniewski, E. J. (1997). When concepts combine. *Psychonomic Bulletin and Review*, 4, 167–183.
- Wisniewski, E. J. (2002). Concepts and categorization. In D. L. Medin (Ed.), *Steven’s handbook of experimental psychology* (3rd ed., pp. 467–532). New York: Wiley.
- Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science*, 18, 221–281.
- Wolff, P., & Song, G. (2003). Models of causation and the semantics of causal verbs. *Cognitive Psychology*, 47, 241–275.
- Wolff, P., Medin, D. L., & Pankratz, C. (1999). Evolution and devolution of folkbiological knowledge. *Cognition*, 73, 177–204.
- Xu, F. (2003). The development of object individuation in infancy. In H. Hayne & J. Fagen (Eds.), *Progress in infancy research*, Vol. 3 (pp. 159–192). Mahwah, NJ: Erlbaum.
- Xu, F., & Carey, S. (1996). Infants’ metaphysics: The case of numerical identity. *Cognitive Psychology*, 30, 111–153.
- Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language*, 39, 124–149.
- Yamauchi, T., & Markman, A. B. (2000a). Inference using categories. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 26, 776–795.
- Yamauchi, T., & Markman, A. B. (2000b). Learning categories composed of varying instances: The effect of classification, inference and structural alignment. *Memory and Cognition*, 28, 64–78.
- Yamauchi, T., Love, B. C., & Markman, A. B. (2002). Learning non-linearly separable categories by inference and classification. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 28, 585–593.
- Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, 81, 141–145.
- Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8, 338–353.

Causal Learning

Patricia W. Cheng *and* Marc J. Buehner

Abstract

This chapter is an introduction to the psychology of causal inference using a computational perspective, with the focus on causal discovery. It explains the nature of the problem of causal discovery and illustrates the goal of the process with everyday and hypothetical examples. It reviews psychological research under two approaches to causal discovery, an associative approach and a causal approach that incorporates causal assumptions in the inference process. The latter approach provides a framework within which to answer different questions regarding causal inference coherently. The chapter ends with a consideration of causality as unfolding over time. We conclude with a sketch of future directions for the field.

Key Words: causal learning, rationality, associative models, causal models, causal Bayes nets, Bayesian causal models, temporal contiguity, causal invariance, intervention, empirical knowledge

Why Causality?

The central question of this chapter is: “How can any intelligent system put on Planet Earth, if given cognitive resources and types of information similar to those available to us, discover how the world works so that it can best achieve its goals?” Before we attempt to answer this question, let us imagine that our cognition were different in various respects. First, suppose we were unable to learn associations between events (i.e., detect statistical regularity in the occurrence of events). We would be unable to predict any events, causal or otherwise. For example, we would be unable to predict that if the traffic light turns red, we should stop or an accident is likely to happen, or that if we hear a knock on our door, someone will be on the other side when we open the door. Nor would we be able to predict the weather, even imperfectly. We would be unable to learn the sequences of sound in language or music, or the meaning of words. We would behave as if we had prosopagnosia, unable to relate our parents’

faces, or the face of the person we have been dating for the past month, to their past history. A nonassociative world would be grim.

Now, imagine a world where we were able to learn associations but unable to reason about causes and effects. What would it be like? In that world, for a child growing up on a farm who has always experienced sunrise after the rooster crows, if the rooster is sick one morning and does not crow, she would predict that the sun would not rise until the rooster crows. Similarly, if the rooster has been deceived into crowing, say, by artificial lighting in the middle of the night, the child would predict that the sun would rise soon after. Notice that under normal conditions noncausal associations do enable one to reliably predict a subsequent event from an observation (e.g., sunrise from a rooster’s crowing, a storm soon to come from a drop in the barometric reading, or from ants migrating uphill). They do not, however, support predictions about the event (e.g., sunrise) when the observation (crowding or no

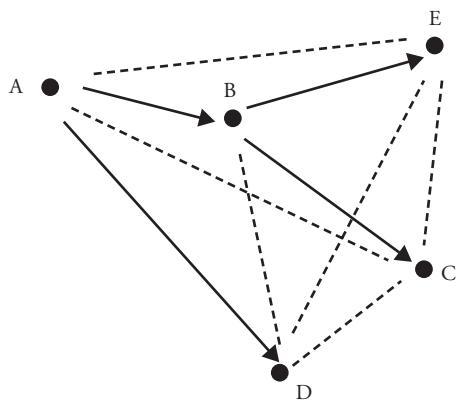


Fig. 12.1 A causal tree and the implied associative links. Nodes represent variables, labeled by letters. Arrows represent direct causal links. Dotted lines represent implied associations, which include indirect causal links as well as associations between direct and indirect effects of a common cause.

crowning) is produced by an action or an extraneous cause (respectively, artificial light and the rooster's sickness). An associative world without causation would be exasperating.

Consider how often we would be wrong, and how inefficient we would be, were we to store all associations, both causal and noncausal. We illustrate the problem with the causal tree in Figure 12.1. In the figure, each node represents a variable, and each arrow represents a causal relation. There are four causal links, but six additional associations (the dotted lines).¹ These additional associations can be inferred from the causal links, and thus are redundant. In an associative world, if information on any of the six extra associations is salient (e.g., as information on a rooster's crowing and sunrise might be), they would be indistinguishable from the causal links. Thus, not only would the extra associations be inefficient to store, many would yield erroneous predictions based on actions. For example, node D in the figure is linked by a single arrow, but by three additional associations, to nodes B, C, and E; manipulating any of these variables would not lead to D, the “desired” outcome predicted by these three associations.

Finally, not only would we be unable to achieve our goals, but we would be unable to structure an otherwise chaotic flux of events into meaningful episodes. We explain events by causation. Returning to our storm example, whereas we might explain that a car skidded and rolled down the mountain-side because of the rainstorm, it would be odd to explain that the car skidded because of the low

barometric reading. Causal explanations are universal, as anthropologists who study everyday narratives across cultures have observed; they serve to imbue life events with an orderliness, to demystify unexpected events, and establish coherence (Ochs & Capps, 2001).

Predicting the Consequences of Actions to Achieve Goals: A Framework for Causal Learning, Category Formation, and Hypothesis Revision

For the just mentioned reasons, it is easy to see why it is important to distinguish between causation and mere association. A less obvious reason, one that has implications for the formulation of the problem to be solved, is that whereas associations are observable, causal relations are inherently unobservable and can only be inferred. For example, one can observe the number of lung cancer patients among cigarette smokers and among nonsmokers and see the association between cigarette smoking and lung cancer, but the association can be causal or noncausal. It may, for example, be due to confounding by the higher incidence of radon in the smokers' dwellings, and the exposure to radon is what caused the smokers' lung cancer. The challenge is how to encapsulate causal relations, even though they are unobservable, so that the causal knowledge can be applied to best achieve desired outcomes.

We have so far implicitly assumed that cause-and-effect variables, such as “sunrise,” “drop in barometric reading,” and “storm,” are predefined, given to the causal learner, and only the relations between them are to be discovered. A more realistic description of the situation is: In order for our causal knowledge to be generalizable from the learning context (e.g., prior experience, whether one's own or that of others, shows that icy roads cause skidding) to the application context (I don't want my car to skid, so I will wait until the ice has melted before I drive), we construct a representation of the world in which cause-and-effect variables are so defined that they enable ideally invariant causal relations to be constructed. As the philosopher C. I. Lewis (1929) observed, “Categories are what obey laws.” Defining the objects, events, and categories linked by causal relations is part of the problem of causal discovery. Fortunately for cognitive psychologists studying human causal discovery, some of the work defining variables is already taken care of by evolution, prewired into our perceptual system and emotions. For example, we see a rooster as figure

against the ground of the farm landscape. Strong gusts of wind alarm us, and getting wet in the rain is unpleasant. But there is definitely remaining work; for example, what determines that “ants migrating uphill” should be defined as a variable?

Whenever we apply causal knowledge to achieve a goal, we are assuming that the causal relations in question remain invariant from the learning context to the application context. Because of our limited causal knowledge, however, a causal relation that we assume to be invariant would no doubt in fact often change (e.g., a scientist might hypothesize “vitamin E has antioxidant effects” but find instead that whereas natural foods rich in vitamin E have antioxidant effects, vitamin E pills do not). The assumption of causal invariance in our everyday application of causal knowledge might seem too strong. Although simplistic as a static hypothesis, however, this assumption is rational as a defeasible default within the dynamic process of hypothesis testing and hypothesis revision. Given our limited access to information at any particular moment, the criterion of causal invariance serves as a compass aimed at formulating the simplest explanation of a phenomenon that allows invariance to obtain (e.g., the scientist might search for other substances in natural foods that in conjunction with vitamin E consistently produce the effects). Observed deviation from the default indicates a need for hypothesis revision, a change in direction aimed at capturing causal invariance (Carroll & Cheng, 2010).

Cheng (2000) showed that an alternative assumption that would also justify generalization of a causal relation regarding an outcome to a new context is that enabling conditions (causal factors in the contextual background interacting with the hypothesized cause) and preventive causes that occur in the background (all of which are often unobserved) occur just as frequently in the generalization context as in the learning context. She also showed that with respect to the accuracy of generalization to new contexts, the two assumptions are equivalent. In the rest of our chapter, we use causal invariance (which we term “independent causal influence” when we define it mathematically) because it is the simpler of the two equivalent conceptions.

Now that we have considered some goals and constraints of causal inference, let us rephrase the question of causal learning with respect to those goals and constraints: How can any intelligent agent given the information and resources available to humans discover ideally invariant causal relations

that support generalization from the learning context to an application context? In particular, would it suffice to have a powerful statistical process that detects regularities among events but lacks any a priori assumptions about how the world works? Because humans have limited access to information, an accompanying question is, What hypothesis testing and revision process would allow the ideally invariant causal relations to be constructed?

By posing our question in terms of discovery, we by no means rule out the possibility that there exist some innate domain-specific biases. Classic studies by Garcia, McGowan, Ervin, and Koelling (1968) demonstrated two such biases. In each of four groups of rats, one of two cues, either a novel size or a novel flavor of food pellets, was conditionally paired with either gastrointestinal malaise induced by X-ray or with pain induced by electrical shock. The combination of flavor and illness produced a conditioned decrement in the amount consumed but that of the size of the pellet and illness did not. Conversely, the combination of size and pain produced hesitation before eating, but flavor and pain did not. Apparently, the novelty had to be of the right kind for effective causal learning regarding the malaise and shock to occur.

Most of what we do know about the world, however, must have been acquired due to experience. How else could we have come to know that exposure to the sun causes tanning in skin but causes bleaching in fabrics? Or come to know that billiard ball A in motion hitting billiard ball B at rest would not jump over B, rebound leaving B still, or explode (Hume, 1739/1888)? Notice that it is not necessary for the causal learner to know *how* sunlight causes tanning in skin or bleaching in fabrics to discover that it does. Neither was it necessary, for that matter, for the rats to know *how* the X-ray or electricity caused their respective effects for their learning to occur. We will return to the issue of adding intervening nodes in a causal network to explain how an outcome is achieved via a causal mechanism.

Proposed Solutions: Two Dominant Approaches

We have only gone so far as posing the problem to be solved. Hopefully, posing the problem clearly will mean much of the work has been done. In the rest of this chapter, we review proposed solutions according to two dominant approaches: the associative approach, including its statistical variants, and the causal approach. We follow each of these

accounts with a review of main empirical tests of the approach. (For a discussion of how the perceptual view [Michotte, 1946/1963], the mechanism view [Ahn, Kalish, Medin, & Gelman, 1995], and the coherence view [Thagard, 2000] relate to these approaches, see Buehner and Cheng [2005].) We then broaden our scope to consider the role of temporal information in causal learning. We end the chapter with a sketch of future research directions.

The Associative Approach

An intuitive approach that has dominated psychological research on learning is the associative approach (e.g., Allan & Jenkins, 1980; Jenkins & Ward, 1965; Rescorla & Wagner, 1972), which traces its roots to the philosopher David Hume (1739/1888). Hume made a distinction between analytic and empirical knowledge, and argued that causal knowledge is empirical. Only experience tells us what effect a cause has. The strong conviction of causality linking two constituent events is but a mental construct. The problem of causal learning posed by Hume radically shaped subsequent research on the topic and set the agenda for the study of causal learning from a cognitive science perspective. Both the associative and causal approaches are predicated on his posing of the problem.

To Hume, the relevant observed aspects of experience that give rise to the mentally constructed causal relations were the repeated association between the observed states of a cause and its effect, their temporal order and contiguity, and spatial proximity. Our examples have illustrated that one can predict a future event from a *covariation*—the concerted variation among events—provided that causes of that event remain unperturbed. Predictions of this kind are clearly useful; we appreciate weather reports, for example. To early associative theorists, causality is nothing more than a fictional epiphenomenon floating unnecessarily on the surface of indisputable facts.² After all, causal relations are unobservable. In fact, Karl Pearson, one of the fathers of modern statistics, subscribed to a positivist view and concluded that calculating correlations is the ultimate and only meaningful transformation of evidence at our disposal: “Beyond such discarded fundamentals as ‘matter’ and ‘force’ lies still another fetish amidst the inscrutable arcana of modern science, namely, the category of cause and effect” (Pearson, 1911, p. iv).

But, as we saw earlier, mere associations are inadequate for predicting the consequences of actions

and would also be inefficient to store. Thus, in addition to dissecting the traditional associative view to understand its shortcomings, we will also consider a more viable augmented variant of the associative view, one similar to how scientists infer causation. The augmented view assumes that rational causal learning requires not only a sophisticated detector of covariations among events but also the use of actions as a causality marker: When the observed states of events are obtained by an action, by oneself or others, intervening in the normal course of events, the observed associations are causal; otherwise, they are noncausal. After all, one can observe that actions are what they are; there is therefore no deviation from Hume’s constraint that causal discovery begins with observable events as input. In entertaining this variant, we are taking the perspective of the design of an intelligent causal learner on our planet, rather than adhering to how the associative view has been traditionally interpreted. This more viable variant of the associative view implicitly underlies the use of associative statistics in typical tests of causal hypotheses in medicine, business, and other fields. It retains the strong appeal of the associative approach, namely, its objectivity. Other things being equal, positing unobservable events, as the causal view does, seems objectionable.

A growing body of research is dedicated to the role of intervention in causal learning, discovery, and reasoning (e.g., Blaisdell, Sawa, Leising, & Waldmann, 2006; Gopnik et al., 2004; Lagnado & Sloman, 2004; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). Indeed, the general pattern reported is that observations based on intervention allow causal inferences that are not possible based on mere observations.

A STATISTICAL MODEL

For situations involving only one varying candidate cause, an influential decision rule for more than four decades has been the ΔP rule:

$$\Delta P = P(e^+|c^+) - P(e^+|c^-) \quad (1)$$

according to which the strength of the relation between binary causes c and effects e is determined by their *contingency* or *probabilistic contrast*—the difference between the probabilities of e in the presence and absence of c (see, e.g., Allan & Jenkins, 1980; Jenkins & Ward, 1965). ΔP is estimated by relative frequencies. In our equations, we denote the

		Effect <i>e</i>	
		present	absent
Candidate cause <i>c</i>	present	A	B
	absent	C	D

Fig. 12.2 A standard 2×2 contingency table; *a* through *d* are labels for event types resulting from factorial combination of the presence and absence of cause *c* and effect *e*.

“presence” value of a binary variable by a “+” superscript and the “absence” value by a “−” superscript (e.g., $c+$ denotes the presence of *c*). Figure 12.2 displays a standard *contingency table* where cells *A* and *B* respectively represent the frequencies of the occurrence, and nonoccurrence, of *e* in the presence of *c*; cells *C* and *D* represent, respectively, the frequencies of the occurrence, and of nonoccurrence, of *e* in the absence of *c*.

If ΔP is noticeably positive, then *c* is thought to produce *e*; if it is noticeably negative, then *c* is thought to prevent *e*; and if ΔP is not noticeably different from zero, then *c* and *e* are thought not to be causally related to each other. Several modifications of the ΔP rule to include various parameters have been proposed (e.g., Anderson & Sheu, 1995; Perales & Shanks, 2007; Schustack & Sternberg, 1981; White, 2002). By allowing extra degrees of freedom, these modified models fit certain aspects of human judgment data better than the original rule. Another type of modification is to compute the ΔP value of a candidate cause conditioned on constant values of alternative causes (Cheng & Holyoak, 1995). This modification allows the model to better account for the influence of alternative causes (as illustrated later). Like all other psychological models of causal learning, all variants of the ΔP model assume that the candidate causes are perceived to occur before the effect in question.

AN ASSOCIATIONIST MODEL

In the domain of animal learning, an organism’s capacity to track contingencies in its environment has long been of central interest, and apparent parallels between conditioning and causal

learning have led many researchers (see Shanks & Dickinson, 1987; for a review see De Houwer & Beckers, 2002) to search for explanations of human causal learning in neural-network models that specify the algorithm of learning. The most influential associationist theory, the Rescorla-Wagner (RW) model (Rescorla & Wagner, 1972), and all its later variants, is based on an algorithm of error correction driven by a discrepancy between the expected and actual outcomes. For each learning trial where a cue was presented the model specifies

$$\Delta V_{CS} = \alpha_{CS} \beta_{US} (\lambda - \Sigma V) \quad (2)$$

where ΔV is the change in the strength of a given CS-US association on a given trial (CS stands for conditioned stimulus, e.g., a tone; US stands for unconditioned stimulus, e.g., a foot-shock), α and β represent learning rate parameters reflecting the saliences of the CS and US, respectively, λ stands for the actual outcome of each trial (usually 1.0 if it is present and 0 if it is absent), and ΣV is the expected outcome defined as the sum of all associative strengths of all CSs present on that trial.

For situations involving only one varying cue, its mean weight at equilibrium according to the RW algorithm has been shown to equal ΔP if the value of β remains the same when the US is present and when it is absent for the λ values just mentioned (Chapman & Robins, 1990; Danks, 2003). In other words, this simple and intuitive algorithm elegantly explains why causal learning is a function of contingency. It also explains a range of results for designs involving multiple cues, such as *blocking* (see section on “Blocking” to follow), *conditioned inhibition*, *overshadowing*, and *cue validity* (Miller, Barnet, & Grahame, 1995).

Blocking: Illustrating an Associationist Explanation

“Blocking” (Kamin, 1969) occurs after a cue (*A*) is established as a perfect predictor (*A+*, with “+” representing the occurrence of the outcome), followed by exposure to a compound consisting of *A* and a new, redundant, cue *B*. If *AB* is also always followed by the outcome (*AB+*), cue *B* receives very little conditioning; its conditioning is *blocked* by cue *A*. According to RW, *A* initially acquires the maximum associative strength supported by the stimulus. Because the association between *A* and the outcome

is already at asymptote when B is introduced, there is no error left for B to explain, hence the lack of conditioning to B. What RW computes is the ΔP for B conditioned on the constant presence of A. Shanks (1985) replicated the same finding in a causal reasoning experiment with human participants, although the human responses seemed to reflect uncertainty of the causal status of cue B rather than certainty that it is noncausal (e.g., Waldmann & Holyoak, 1992).

Failure of the RW Algorithm to Track Covariation When a Cue Is Absent

However, Shanks' (1985) results also revealed evidence for *backward blocking*; in fact, there is evidence for backward blocking even in young children (Gopnik et al., 2004). In this procedure, the order of learning phases is simply reversed; participants first learn about the perfect relation between AB and the outcome (AB+), and subsequently learn that A by itself is also a perfect predictor (A+). Conceptually, forward and backward blocking are identical, at least from a causal perspective. A causal explanation might go: If one knows that A and B together always produce an effect, and one also knows that A by itself also always produces the effect, one can infer that A is a strong cause. B, however, might be a cause, even a strong one, or noncausal; its causal status is unclear. Typically, participants express such uncertainty with low to medium ratings relative to ratings for control cues that have been paired with the effect an equal number of times (see Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008, for a review).

Beyond increasing susceptibility to attention and memory biases (primacy and recency; see, e.g., Dennis & Ahn, 2001), there is no reason why the temporal order in which knowledge about AB and A is acquired should play a role from a rational standpoint. This is not so for the RW model, however. The model assumes that the strength of a cue can only be updated when that cue is present. In the backward blocking paradigm, however, participants *retrospectively* alter their estimate of B on the A+ trials in phase 2. In other words, the ΔP of B, conditioned on the presence of A, decreases over a course of trials in which B is actually absent, and the algorithm therefore fails to track its covariation.

Several modifications of RW have been proposed to allow the strengths of absent cues to be changed, for instance, by setting the learning parameter α negative on trials where the cue is absent: Van Hamme and Wasserman's (1994) modified RW model, Dickinson and Burke's modified sometimes-opponent-process

model (1996), and the comparator hypothesis (Denniston, Savastano, & Miller, 2001; Miller & Matzel, 1988; Stout & Miller, 2007). Such modifications can explain backward blocking and some other findings showing retrospective revaluation (for an extensive review of modifications to associative learning models applicable to human learning see De Houwer & Beckers, 2002). But these modifications also oddly predict that one will have difficulty learning that there are multiple sufficient causes of an effect. For example, if one drinks tea by itself and finds it quenching, but one sometimes drinks both tea and lemonade, then learning subsequently that lemonade alone can quench thirst will cause one to unlearn that tea can quench thirst. Carroll, Cheng, and Lu (2010) found that in such situations human subjects do not revise causal relations for which they have unambiguous evidence (e.g., that tea is quenching).

Causal Inference: Empirical Findings on Humans and Rats

Association does not equal causation, as we illustrated earlier and as every introductory statistics text warns. We now review how humans and rats reason causally rather than merely associatively.

THE DIRECTION OF CAUSALITY

The concept of causality is fundamentally directional (Reichenbach, 1956) in that causes produce effects, but effects cannot produce causes. Thus, whereas we might say that, given the angle of the sun at a certain time of the day, the height of a flagpole explains the length of its shadow on the ground, it would be odd to say the reverse.³ A straightforward demonstration that humans make use of the direction of the causal arrow was provided by Waldmann and Holyoak (1992), who reasoned that only causes, but not effects, should "compete" for explanatory power. If P is a perfect cause of an outcome A, and R, a redundant cue, is only presented preceding A in conjunction with P, one has no basis of knowing to what extent, if at all, R actually produces A. Consequently, the predictiveness of R should be depressed relative to P in a predictive situation. But if P is instead a consistent effect of A, there is no reason why R cannot also be an equally consistent effect of A. Alternative causes need to be kept constant to allow causal inference, but alternative effects do not. Consequently, the predictiveness of R should not be depressed in a diagnostic situation.

This asymmetry prediction was tested with the blocking design, using scenarios to manipulate whether a variable is interpreted as a candidate

cause or as an effect. Participants in Waldmann and Holyoak's (1992) Experiment 3 had to learn the relation between several light buttons and the state of an alarm system. The instructions introduced the buttons as causes for the alarm in the *predictive* condition, but as potential consequences of the state of the alarm system in the *diagnostic* condition.

Waldmann and Holyoak found exactly the pattern of results they predicted: There was blocking in the predictive condition, but not the diagnostic condition. These results reveal that humans make use of the direction of the causal arrow. Follow-up work from Waldmann's lab (Waldmann & Holyoak, 1997; Waldmann, 2000, 2001) as well as others (Booth & Buehner, 2007; López, Cobos, & Caño, 2005) has demonstrated that the asymmetry in cue competition is indeed a robust finding.

CEILING EFFECTS

One might think that augmenting statistical models with intervention would solve the problem of the directionality of causation. But although intervention generally allows causal inference, it does not guarantee it. Consider a food allergy test that introduces samples of food into the body by needle punctures on the skin. The patient may react with hives on all punctured spots, and yet one may not know whether the patient is allergic to any of the foods. Suppose her skin is allergic to needle punctures, so that hives appear also on punctured spots without food. In this example, there is an intervention, but no causal inference regarding food allergy seems warranted (Cheng, 1997). More generally, interventions are subject to the problem of the well-known *placebo effect*, in which the intended intervention is accompanied by a concurrent intervention (as adding allergens into the bloodstream is accompanied by the puncturing the skin), resulting in confounding. Our example illustrates that intervention does not guarantee causal inference. Not only is intervention insufficient for differentiating causation from association, it is also unnecessary. Mariners since ancient times have known that the position and phase of the moon is associated with the rising and falling of the tides (Salmon, 1989). Notably, they did not consider the association causal, and they had no explanation for the ebb and flow of the tides, until Newton proposed his law of universal gravitation. No intervention on the moon and the tides is possible, but there was nonetheless a dramatic change in causal assessment.

A revealing case of the distinction between covariation and causation that does not involve confounding has to do with what is known in experimental design as a *ceiling effect*. We illustrate this effect with the preventive version of it (a principle never covered in courses on experimental design); the underlying intuition is so powerful it needs no instruction. Imagine that a scientist conducts an experiment to find out whether a new allergy drug relieves migraine as a side effect. She follows the usual procedure and administers the medicine to an experimental group of patients, while an equivalent control group receives a placebo. At the end of the study, the scientist discovers that none of the patients in the experimental group but also none of the patients in the control group suffered from migraine. The effect never occurred, regardless of the intervention. If we enter this information into the ΔP rule, we see that $P(e^*|c^*) = 0$ and $P(e^*|c) = 0$, yielding $\Delta P = 0$. According to the ΔP rule and RW, this would indicate that there is no causal relation, that is, the drug does not relieve migraine. Would the scientist really conclude that? No, the scientist would instead recognize that she has conducted a poor experiment and hence withhold judgment on whether the drug relieve migraine. If the effect never occurs in the first place, how can a preventive intervention be expected to prove its effectiveness?

Even rats seem to appreciate this argument (Zimmer-Hart & Rescorla, 1974). For associative models, however, when an inhibitory cue (i.e., one with negative associative strength) is repeatedly presented without the outcome, so that the actual outcome is 0 whereas the expected outcome is negative, the prediction is that the cue reduces its strength toward 0. That is, in a noncausal world, we would unlearn our preventive causes whenever they are not accompanied by a generative cause. For example, if we inoculate child after child with polio vaccine in a country, and there is no occurrence of polio in that country, we would come to believe that the polio vaccine does not function anymore, rather than merely that it is not needed. To the contrary, even for rats, the inhibitory cue retains its negative strength (Zimmer-Hart & Rescorla, 1974). In other words, when an outcome in question never occurs, either when a conditioned inhibitory cue is present or when it is not, rats apparently treat the zero ΔP value as uninformative, retaining the inhibitory status of the cue. In this case, in spite of a discrepancy between the expected and actual outcomes, there is no revision of causal strength.

Table 12.1. Relative Frequencies of Headache and Model Values for Each Hypothetical Study and Each Condition in Liljeholm and Cheng (2007, Experiment 2)

	Study 1		Study 2		Study 3	
	e no A	e A	e no A	e A	e no A	e A
Varying-base rate	16/24	22/24	8/24	20/24	0/24	18/24
Constant base rate	0/24	6/24	0/24	12/24	0/24	18/24

Notice that given the aforementioned hypothetical migraine-relief experiment, from the same exact data, showing that migraine never occurs one can conclude that the drug *does not cause* migraine rather than withhold judgment. Thus, given the exact same covariation, the causal learner can simultaneously have two conclusions depending on the direction of influence under evaluation (generative vs. preventive). Wu and Cheng (1999) conducted an experiment that showed that beginning college students, just like experienced scientists, do and do not refrain from making causal inferences in the generative and preventive ceiling effects situations depending (in opposite ways) on the direction of influence to be evaluated. We are not aware of any convincing modification of associationist models that can accommodate the finding.

DEFINITION OF CAUSAL INVARIANCE: BEYOND AUGMENTATION OF ASSOCIATIONS WITH INTERVENTION AND OTHER PRINCIPLES OF EXPERIMENTAL DESIGN

The same problem that leads to the ceiling effect—namely, the lack of representation of causal relations—manifests itself even when *all* the principles of experimental design are obeyed. Even in that case, the associative view makes anomalous predictions. Liljeholm and Cheng (2007, Experiment 2) presented college students with a scenario involving three studies of a single specific cue A (Medicine A, an allergy medicine) as a potential cause of an outcome (headache as a potential side-effect of the allergy medicine). In the scenario, allergy patients in the studies were randomly assigned to an experimental group that received Medicine A and a control group that received a placebo. In the three studies, the probability of the outcome was higher by, respectively, $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{3}{4}$ in the experimental group than in the control group (i.e., $\Delta P = \frac{1}{4}$, $\frac{1}{2}$, and $\frac{3}{4}$; see Table 12.1). In a varying-base-rate condition, the base rate of headache differed across the three studies. In a constant-base-rate condition, the

base rate of the effect remained constant: Headache never occurred without the medicine. The students were asked to assess whether the medicine interacted with unobserved causes in the background across the studies or influenced headache the same way across them. As intuition suggests, more students in the constant-base-rate condition than in the varying-base-rate condition (13 out of 15, vs. 5 out of 15, respectively) judged the medicine to interact with the background causes.

Because the changes in covariation, as measured by associative models such as ΔP (Jenkins & Ward, 1965) or RW (Rescorla & Wagner, 1972), were the same across conditions, these associative models could not explain the observed pattern of judgments. Thus, even when there is an effective intervention and there is no violation of the principles of experimental design, a statistical account will not suffice. We return to discuss the implications of these results later.

INTERVENTION VERSUS OBSERVATION

Following analogous work on humans (Waldmann & Hagmayer, 2005), Blaisdell et al. (2006) reported a result that challenges associative models: Rats are capable of distinguishing between observations and interventions. In Experiment 1, during a Pavlovian learning phase rats were trained on two interspersed pairs of associations: A light cue (L) is repeatedly followed by either a tone (T) or food (F). If the rats learned that L is a common cause of T and F (see Fig. 12.5a), then in the test phase, observing T should lead them to infer that L must have occurred (because L was the only cause of T), which should in turn lead them to predict F (because L causes F). The number of nose pokes into the food bin measures prediction of F. Consider an alternative condition in which during a test phase the rats learn that pressing on a newly introduced lever turns on T. Because generating T by means of an alternative cause does not influence its cause (L), turning T on by pressing a lever should not lead the rats to

predict F. After the learning phase, rats were allocated to either the observation or the intervention condition. The occurrences of T in the test phase were equated across the two conditions by yoking the observation rats to the intervention rats, so that when a rat in the intervention condition pressed the lever and T followed, a rat in the observation condition heard T at the same time, independently of their lever pressing. L and F never occurred during the test phase. Remarkably, the observation rats nose-poked more often than the intervention rats in the interval following T, even though during the learning phase, T and F never occurred simultaneously on the same trial.

Because all occurrences of L, T, and F were identical across the observation and intervention groups, associations alone cannot explain the difference between observing T and intervening to obtain T. Even if augmented with the assumption that interventions have special status, so that the pairing between lever pressing and T, for example, is learned at a much faster rate than purely observed pairings, there would still be no explanation for why the intervention rats apparently associate T with L less than did the observation rats. We will return to discuss a causal account of the observed difference.

A Causal Approach

A solution to the puzzles posed by the distinction between covariation and causation is to have a leap of faith that causal relations exist, even though they are unobservable (Kant, 1781/1965). This leap of faith distinguishes the diverse variants of the causal approach from all variants of the associative approach. Some psychologists have proposed that human causal learning involves positing candidate causal relations and using deductive propositional reasoning to arrive at possible explanations of observed data (De Houwer, Beckers, & Vandorpe, 2005; Lovibond, Been, Mitchell, Bouton, & Frohardt, 2003; Mitchell, De Houwer, & Lovibond, 2009). Others (Gopnik et al., 2004) have proposed that human causal learning is described by *causal Bayes nets*, a formal framework in which causal structures are represented as directed acyclic graphs (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993/2000; see Sloman, 2005, for a more accessible exposition). The graphs consist of arrows connecting some nodes to other nodes, where the nodes represent variables and each arrow represents a direct causal relation between two variables;

“acyclic” refers to the constraint that the paths formed by the arrows are never loops. Others have proposed a variant of causal Bayes nets that makes stronger causal assumptions; for example, assume as a defeasible default that causes do not interact, and revise that assumption only when there is evidence against it. The stronger assumptions enable the learner to construct causal knowledge incrementally (Buehner, Cheng, & Clifford, 2003; Cheng, 1997, 2000; Griffiths & Tenenbaum, 2005; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008; Waldmann, Cheng, Hagmayer, & Blaisdell, 2008).

These variants of the causal view, in addition to their explicit representation of causal relations, share a rational perspective (see Chater & Oaksford, Chapter 2). Thus, they all have a goal of inferring causal relations that best explain observed data. They all make use of deductive inference (for examples of the role of analytic reasoning in empirical learning, see Mermin, 2005; Shepard, 2008). It may be said that they all assume that the causal learner deduces when to induce! Our focus in this chapter is on explaining basic ways in which the causal approach provides a solution to what appears to be impasses from an associative perspective.

A THEORY OF CAUSAL INDUCTION

We use Cheng (1997)’s *causal power* theory (also called the power PC theory, short for “a causal power theory of the probabilistic contrast model”) to illustrate how a causal theory explains many of the puzzles mentioned earlier. This theory starts with the Humean constraint that causality can only be inferred, using observable evidence (e.g., covariations, temporal ordering, and spatial information) as input to the reasoning process. It combines that constraint with Kant’s (1781/1965) postulate that reasoners have a priori notions that types of causal relations exist in the universe.

This unification can best be illustrated with an analogy. The relation between a causal relation and a covariation is like the relation between a scientific theory and a model. Scientists postulate theories (involving unobservable entities) to explain models (i.e., observed regularities or laws); the kinetic theory of gases, for instance, is used to explain Boyle’s law. Boyle’s law describes an observable phenomenon, namely that $\text{pressure} \times \text{volume} = \text{constant}$ (under certain boundary conditions), and the kinetic theory of gases explains in terms of unobservable entities why Boyle’s law holds (gases consist of small particles moving at a speed proportional to their temperature,

and pressure is generated by the particles colliding with the walls of the container). Likewise, a causal relation is the unobservable entity that reasoners strive to infer in order to explain observable regularities between events. This distinction between a causal relation as an unobserved, distal, postulated entity and covariation as an observable, proximal stimulus property implies that there can be situations where evidence is observable, but inference is not licensed, and the goal of causal inference thus cannot be met. Specifically, this means that the desired distal unknown, such as causal strength, is represented as a variable (cf. Doumas & Hummel, Chapter 4; Holyoak & Hummel, 2000), separately from covariation, allowing situations where covariation has a definite value (e.g., 0, as in the ceiling effect), but the causal variable has no value.

How, then, does the causal power theory (Cheng, 1997) go beyond the proximal stimulus and explain the various ways in which covariation does not imply causation? The path through the derivation of the estimation of causal strength reveals the answers. For inferring simple (i.e., elemental) causal relations, the theory partitions all causes of effect e into the candidate cause in question, c , and a , a composite of all (observed and unobserved) alternative causes of e . “Alternative causes” of e include all and only those causes of e that are not on the same causal path to e as c . Thus, c can be thought of as a composite that includes all causes on the same causal path as c preceding e . This partitioning is a general structure that maps onto all learning situations involving candidate causes and effects that are binary variables with a “present” and an “absent” value. We focus on this type of variables because they best reveal how the associative and causal views differ.

The unobservable probability with which c produces e (i.e., the probability that e occurs as a result of c occurring), termed the *generative* causal power of c with respect to e , is represented by a variable, q_c . When $\Delta P \geq 0$, q_c is the desired unknown. Likewise, when $\Delta P \leq 0$, the *preventive* causal power of c , denoted by p_c , is the desired unknown. Two other relevant theoretical unknowns are q_a , the probability with which a produces e when it occurs, and $P(a)$, the probability with which a occurs. The composite a may include unknown or unobservable causes. Because any causal power variable may have a value of 0, or an unknown or undefined value, these variables are merely hypotheses—they do not presuppose that c and a indeed have causal influence on e . The idea of a cause producing an effect and of a cause preventing an effect are

primitives in the theory (see Goodman, Ullman, & Tenenbaum, 2011, and Tenenbaum, Kemp, Griffiths & Goodman, 2011, for an alternative view).

The theory assumes four general simplifying beliefs (Cheng, 1997; Novick & Cheng, 2004):

1) c and a influence e independently,

2) a could produce e but not prevent it,

3) causal powers are independent of the

frequency of occurrences of the causes (e.g., the causal power of c is independent of the frequency of occurrence of c), and

4) e does not occur unless it is caused.

Assumption 1 is a leap of faith inherent to this incremental learning variant of causal discovery. This is the defeasible default assumption we termed “causal invariance” earlier. Two causes influencing effect e “independently” means that the influence of each on e remains unchanged regardless of whether e is influenced by the other cause. Assumption 2 is likewise a default hypothesis, adopted unless evidence discredits it. (Alternative models apply if assumption 1 or 2 is discredited; see Novick & Cheng 2004; see Cheng, 2000, for implications of the relaxation of these assumptions.) This set of assumptions, which is stronger than that assumed in standard Bayes nets, enables causal relations to be learned one at a time, when there is information on only the occurrences of two variables, a single candidate cause and an effect. The type of learning described by the theory therefore requires less processing capacity. It is assumed that, as in associative models, when there is information on which variable is an effect, the causal learner iterates through candidate causes of the effect, grouping all potential causes other than the candidate in question as the composite alternative cause. Otherwise, the causal learner iterates through all possible variable pairs of candidate causes and effects.

These assumptions imply a specific function for integrating the influences of multiple causes (Cheng, 1997; Glymour, 2001), different from the additive function assumed by associative models. For the situation in which a potentially generative candidate cause c occurs independently of other causes, the probability of observing the effect e is given by a noisy-OR function,

$$P(e^+ | c; w_a, q_c) = q_c \cdot c + w_a - q_c \cdot c \cdot w_a \quad (3)$$

where $c \in \{0,1\}$ denotes the absence and the presence of the candidate cause c . Recall that in our

equations we denote the “presence” value of a binary variable by a “+” superscript and the “absence” value by a “−” superscript. In contrast, variables have no superscripts. As just mentioned, variable q_c represents the generative power of the candidate cause c . Because it is not possible to estimate the causal power of unobserved causes, variable w_a represents $P(a^+) \cdot q_a$. In the preventive case, the same assumptions are made except that c is potentially preventive. The resulting noisy-AND-NOT integration function for preventive causes is

$$P(e^+ | c; w^a, p_c) = w_a (1 - p_c \cdot c), \quad (4)$$

where p_c is the preventive causal power of c .

Using these “noisy-logical” integration functions (terminology due to Yuille & Lu 2008), Cheng (1997) derived normative quantitative predictions for judgments of causal strength. Under the aforementioned set of assumptions, the causal power theory explains the two conditional probabilities defining ΔP as follows:

$$P(e^+ | c^+) = q_c + P(a^+ | c^+) \cdot q_a - q_c \cdot P(a^+ | c^+) \cdot q_a \quad (5)$$

$$P(e^+ | c^-) = P(a^+ | c^-) \cdot q_a \quad (6)$$

Equation 5 “explains” that given that c has occurred, e is produced by c or by the composite a , nonexclusively (e is jointly produced by both with a probability that follows from the independent influence of c and a on e). Equation 6 “explains” that given that c did not occur, e is produced by a alone.

Explaining the Role of “No Confounding” and Why Manipulation Encourages Causal Inference But Does Not Guarantee Success

It follows from Equations 3 and 4 that

$$\Delta P_c = q_c + P(a^+ | c^+) \cdot q_a - q_c \cdot P(a^+ | c^+) \cdot q_a - P(a^+ | c^-) \cdot q_a \quad (7)$$

From Equation 7, it can be seen that there are four unknowns: q_c , q_a , $P(a^+ | c^+)$, and $P(a^+ | c^-)$! It follows that in general, despite ΔP having a definite value, there is no unique solution for q_c . This failure to solve for q_c corresponds to our intuition that covariation need not imply causation.

When there is no confounding. Now, in the special case in which a occurs independently of c (e.g., when alternative causes are held constant), $P(a^+ | c^+) = P(a^+ | c^-)$. If one is willing to assume “no confounding,” then making use of Equation 6, Equation 7 simplifies to Equation 8,

$$q_c = \frac{\Delta P}{1 - P(e^+ | c^-)} \quad (8)$$

in which all variables besides q_c are observable. In this case, q_c can be solved. Being able to solve for q_c only under the condition of *independent occurrence* explains why manipulation by free will encourages causal inference in everyday reasoning—alternative causes are unlikely to covary with one’s decision to manipulate. For the same reason, it explains the role of the principle of *control* in experimental design.

At the same time, the necessity of the “*no confounding*” condition explains why causal inferences resulting from interventions are not always correct; although alternative causes are unlikely to covary with one’s decision to manipulate, they still may do so, as our needle-puncture allergy example illustrates. Note that the “*no confounding*” condition is a *result* in this theory, rather than an unexplained axiomatic assumption, as it is in current scientific methodology (also see Dunbar & Klahr, Chapter 35).

An analogous explanation yields p_c , the power of c to prevent e

$$p_c = \frac{-\Delta P}{P(e^+ | c^-)} \quad (9)$$

Griffiths and Tenenbaum (2005) showed that if one represents uncertainty about the estimates of causal power by a distribution of the likelihood of each possible strength given the data, then Equation 8 and 9, respectively, are maximum likelihood point estimates of the generative and preventive powers of the candidate cause; that is, they are the peak of the posterior likelihood distributions.

Explaining Two Ceiling Effects

Equations 8 and 9 explain why ceiling effects block causal inference and do so under different conditions for evaluating generative and preventive causal influence. When the outcome does not occur at either a ceiling (i.e., extreme) level, both equations yield either causal power of 0 when the occurrence of c makes no difference to the occurrence of e ($\Delta P = 0$). When e always occurs (i.e., $P(e^+ | c^+) = P(e^+ | c^-) = 1$) regardless

of the manipulation, however, q_e in Equation 8 (the generative case) is left with an undefined value. In contrast, in the preventive case, when e never occurs (i.e., $P(e^+|c^+) = P(e^+|c^-) = 0$), again regardless of the manipulation, p_e in Equation 9 is left with an undefined value.⁴

As we mentioned, most causes are complex, involving not just a single factor but a conjunction of factors operating in concert, and the assumption that c and a influence e independently may be false most of the time. When this assumption is violated, if an alternative cause (part of a) is observable, the independent influence assumption can be given up for the observable alternative cause, and progressively more complex causes can be evaluated using the same distal approach that represents causal powers (see Novick & Cheng, 2004, for an extension of this approach to evaluate conjunctive causes involving two factors). Even if alternative causes are unknown, however, Cheng (2000) showed that as long as they occur with about the same probability in the learning context as in the generalization context, predictions according to simple causal power involving a single factor will hold.

Some have claimed that the causal power approach cannot account for reasoning that combines observations with interventions. As just shown, however, this approach explains the role of interventions in causal learning and how it differs from observation. Likewise indicating that this approach readily accommodates the combination, Waldmann et al. (2008) derived an equation under the causal power assumptions that explains Blaisdell et al.'s results regarding the distinction between observations and interventions in diagnostic reasoning.

Experimental Tests of a Causal Approach

We examine three findings in support of the causal approach. None of these findings can be explained by the associative view, even when augmented with the assumption that only interventions enable causal inference. The first two findings test the two leaps of faith: that causal relations exist and that they are invariant across contexts. The first finding concerns the independent causal influence assumption as manifested in a qualitative pattern of the influence of $P(e^+|c^+)$, the base rate of e , for candidate causes with the same ΔP . The second illustrates the parsimony of a causal explanation that assumes independent causal influence across different types of "effect" variables, specifically, dichotomous and continuous (Beckers, De Houwer, Pineno, & Miller,

2005; Beckers, Miller, De Houwer, & Urushihara, 2006; Lovibond et al., 2003). The third concerns the test reviewed earlier of the distinction between observation and intervention ("seeing" vs. "doing") in diagnostic causal inference (Blaisdell et al., 2006; Waldmann & Hagmayer, 2005). We consider explanations of this distinction as an illustration of the compositionality of the causal view and of the role of deductive reasoning in causal inference.

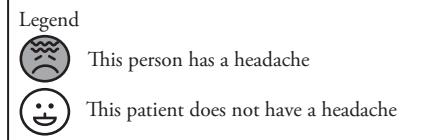
THE INDEPENDENT CAUSAL INFLUENCE ASSUMPTION AFFECTS CAUSAL JUDGMENTS: BASE-RATE INFLUENCE ON CONDITIONS WITH IDENTICAL ΔP

As we noted, a major purpose of causal discovery is to apply the acquired causal knowledge to achieve goals, and that the independent causal influence assumption is a leap of faith that justifies generalization from the learning context to the application context. Here, we see that the assumption leads to causal judgments that differ from those predicted by associative models, even those augmented with a privileged status for interventions and other principles of experimental design. In other words, this assumption not only affects the application of causal knowledge, it affects the very discovery of that knowledge itself.

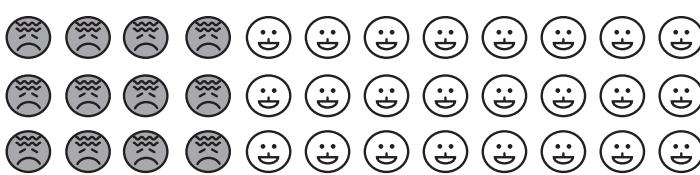
Do people have this leap of faith? Let us examine the predictions based on the causal power assumptions in greater detail. If we consider Equation 8, for any constant positive ΔP , generative causal ratings should *increase* as $P(e^+|c^+)$ increases. Conversely, according to Equation 9, for any constant negative ΔP , preventive causal ratings should *decrease* as $P(e^+|c^-)$ increases. On the other hand, according to both equations, zero contingencies should be judged as noncausal regardless of the base rate of e except when the base rate is at the respective ceiling levels.

No associative model of causal inference, descriptive or prescriptive, predicts this qualitative pattern of the influence of the base rate of e . Normative models are symmetric around the probability of .5 and therefore do not predict an asymmetric pattern either for generative causes alone or for preventive causes alone. Although some psychological associative learning models can explain one or another part of this pattern given felicitous parameter values, the same parameter values will predict notable deviations from the rest of the pattern. For example, in the RW, if $\beta_{us} > \beta_{-us}$ causal ratings for generative and preventive causes will *both* increase as base-rate increases, whereas they will *both* decrease as base-rate

Results for
side-effect of
medicine: B



These patients did not receive Medicine B:



These patients received Medicine B:

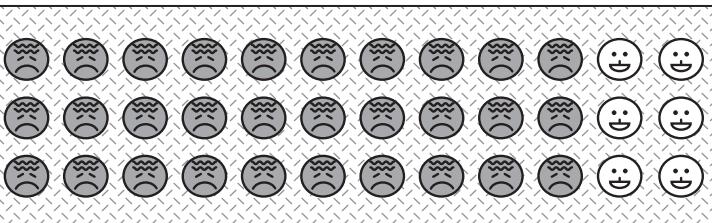


Fig. 12.3 Example stimulus materials from a condition in Buehner et al. (2003).

increases if the parameter ordering was reversed. No consistent parameter setting will predict opposite trends for generative as for preventive causes for the same change in the base rate of e . Another prominent associative learning model, Pearce's (1987) model of stimulus generalization, can account for opposite base rate influences in positive and negative contingencies if the parameters are set accordingly, but this model would then predict a base-rate influence on noncontingent conditions.

Figure 12.3 illustrates the intuitiveness of a deviation from ΔP . The reasoning is counterfactual. $P(e^+|c^-)$, 1/3 in the figure, estimates the "expected" probability of e in the presence of c , if c had been absent so that only causes other than c exerted an influence on e . A deviation from this counterfactual probability is evidence for c being a simple cause of e . Under the assumption that the patients represented in the figure were randomly assigned to the two groups, one that received the drug and another that did not, one would reason that about 1/3 of the patients in the "drug" group would be expected to have

headache if they had not received the drug. For the remaining patients—the 2/3 who did not have already have headaches caused by other factors—the drug would be the sole cause of headaches. In this subgroup, headache occurred in 3/4 of the patients. One might therefore reason, one's best guess for the probability of the drug producing headache is 3/4. If one assumes that for every patient in the control group, regardless of whether the patient had a headache, the drug causes headache with a probability of 3/4, this estimate would result. Among those who already had a headache produced by alternative causes, headache due to the drug is not observable.

In contrast, consider what estimate would result if one assumes instead that the drug causes headache with a probability of 1/2, the estimated causal strength according to associative models such as the ΔP model. Applying that probability to every patient, one's best guess would be that 2/3 of the patients would have headaches after receiving the medicine, rather than the 5/6 shown in Figure 12.3. As should be clear, associative models give estimates

that are inconsistent with the assumption that the causes involved influenced headache independently, even though the additivity in those models is generally assumed to represent independence, and thus to justify generalization to new contexts. This inconsistency, due to the outcome variable being dichotomous, leads to irrational applications of causal knowledge to achieve desired outcomes.

Are people rational or irrational in their estimation of causal strength? To discriminate between the causal power theory and the associative approach, Buehner, Cheng, and Clifford (2003, Experiment 2) made use of the pattern of causal-strength predictions according to Equations 8 and 9 just discussed. They gave subjects a task of assessing whether various allergy medicines have a side effect on headaches, potentially causing headaches or preventing them, when presented with fictitious results of studies on allergy patients (see Fig. 12.3 for an example) in which the patients were randomly assigned to two groups, one receiving the medicine and the other not. The subjects were also asked to rate the causal strengths of each candidate after viewing the results for each fictitious study, using a frequentist counterfactual causal question that specified a novel transfer context: “Imagine 100 patients who do *not* suffer from headaches. How many would have headaches if given the medicine?” The novel context for assessing generative causal power, as just illustrated, is one in which there are no alternative generative causes of headache. By varying the base rate of the target effect, for both generative and preventive causes, the experiment manipulated (1) causal power while keeping ΔP constant, (2) ΔP while keeping causal power constant. The experiment also manipulated the base rate of e for noncontingent candidate causes. Their results clearly indicate that people make the leaps of faith assumed by the causal power theory, contrary to the predictions of all associative models, including normative associative models.

INTEGRATING CAUSAL REPRESENTATION WITH BAYESIAN INFERENCE: REPRESENTING UNCERTAINTY AND EVALUATING CAUSAL STRUCTURE

The reader might have noticed that, just like the ΔP rule, the point estimate of causal power in the causal power theory (Equations 8 and 9) is insensitive to sample size. As initially formulated, the theory did not provide any general account of how uncertainty impacts causal judgments. The point estimates are the most likely strength of the causal link that would have produced the observed data, but causal links with other

strength values, although less likely to have produced the data, could well have also, for smaller sample sizes more so than for larger sizes. The lack of an account of uncertainty in early models of human causal learning, together with methodological problems in some initial experiments testing the causal power theory (see Buehner et al., 2003), contributed to prolonging the debate between proponents of associationist treatments and of the causal power theory. For some data sets (e.g., Buehner & Cheng, 1997; Lober & Shanks, 2000), human causal-strength judgments for some conditions were found to lie intermediate between the values predicted by causal power versus ΔP . This pattern was especially salient for studies in which the causal question, which was ambiguously worded, could be interpreted to concern confidence in the existence of a causal link. Intriguingly, a subtle, statistically insignificant, but consistent trend toward this pattern seemed to occur even for the disambiguated counterfactual question illustrated earlier. These deviations from the predictions of the causal power theory perhaps reflect the role of uncertainty, which is outside the scope of the theory.

An important methodological advance in the past decade is to apply powerful Bayesian probabilistic inference to causal graphs to explain psychological results (e.g., Griffiths & Tenenbaum, 2005; Lu et al., 2008; Tenenbaum et al., 2011; see Griffiths, Tenenbaum, & Kemp, Chapter 3; for a review of recent work, see Holyoak & Cheng, 2011). This new tool enables rationality in causal inference to be addressed more fully. For example, it enables a rich representation of uncertainty and a formulation of qualitative queries regarding causal structure.

Griffiths and Tenenbaum (2005; Tenenbaum & Griffiths, 2001) proposed the *causal support* model, a Bayesian model that addresses the causal query, termed a “structure” judgment, which aims to answer, “How likely is it that a causal link exists between these two variables?” This is in contrast to the causal query regarding causal strength that has been emphasized in previous psychological research on causal learning. Strength judgment concerns the weight on a causal link, which aims to answer the query, “What is the probability with which a cause produces (alternatively, prevents) an effect?”

In terms of the graphs in Figure 12.4, the causal support model aims to account for judgments as to whether a set of observations (D) was generated by Graph 1, a causal structure in which a link may exist between candidate cause C and effect E or by a causal structure in which no link exists between C and E .

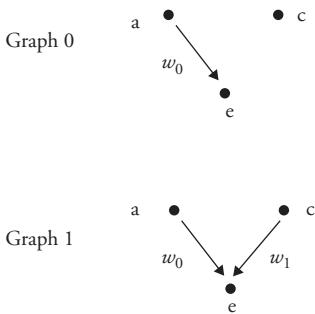


Fig. 12.4 Candidate causal structures varying in whether c causes e .

(Graph 0). Causal-strength models, by contrast, aim to account for people's best estimates of the weight w_1 on the link from C to effect E in Graph 1 that generated D , with w_1 ranging from 0 to 1.

In the causal support model, the decision variable is based on the posterior probability ratio of Graphs 1 and 0 by applying Bayes' rule. Support is defined as:

$$\text{support} = \log \frac{P(D | \text{Graph1})}{P(D | \text{Graph0})}. \quad (10)$$

ASSOCIATIVE VERSUS CAUSAL BAYESIAN MODELS: UNIFORM VERSUS SPARSE AND STRONG PRIORS

Note that adopting Bayesian inference is entirely orthogonal to the longstanding debate between causal and associationist approaches. Because mathematics is a tool rather than an empirical theory, the Bayesian approach can be causal or associative depending on whether causal assumptions are made, even while they are applied to supposedly causal graphs. As Griffiths and Tenenbaum (2005) had noted, a Bayesian model can incorporate either the noisy-logical integration functions derived from the causal power theory or the linear function underlying the Rescorla-Wagner model and the ΔP rule. In addition, a Bayesian analysis can be applied to both strength and structure judgments, as well as to other types of causal queries, such as causal attribution. For strength judgments, rather than basing predictions on the peak of the posterior distribution of w_1 in Graph 1, which corresponds to causal power and ΔP according to the respective models, a natural Bayesian extension of the causal power theory would base predictions on other functions of the posterior distribution of w_1 , such as its mean. Thus, for the fictitious data regarding the side

effect of an allergy medicine in Figure 12.3, rather than estimating that the medicine causes headache with a probability of $3/4$ or $1/2$, as predicted by the causal and associative causal-strength models, respectively, the estimate would fall slightly below $3/4$, in between those estimates.

Lu et al. (2008) developed and compared several variants of Bayesian models as accounts of human judgments about both strength and structure. In addition to directly comparing predictions based on these alternatives, Lu et al. considered two different sets of priors on causal strength. One possible prior is simply a uniform distribution, as assumed in the causal support model. The alternative "generic" (i.e., domain-general) prior tested by Lu et al. is based on the assumption that people prefer parsimonious causal models (Chater & Vitányi, 2003; Lombrozo, 2007; Novick & Cheng, 2004). Sparse and strong (SS) priors imply that people prefer causal models that minimize the number of causes of a particular polarity (generative or preventive) while maximizing the strength of each individual cause that is in fact potent (i.e., of nonzero strength).

The sparse and strong priors, although admitted post hoc, point to the role of parsimony in explanation, an interesting issue for future research. When one is presented with a Necker cube, for example, one perceives two possible orientations. The human perceptual system has implicitly screened out the infinitely many other possible non-cube-shaped objects that would project the same eight corners onto our retina. The visual system makes a parsimony assumption: It favors the simplest "explanations" of the input on our retina. The human causal learning appears to similarly favor parsimonious causal explanations.

For all four Bayesian models, Lu et al. (2008) compared the average observed human strength rating for a given contingency condition with the mean of w_1 computed using the posterior distribution. Model fits revealed that the two causal variants based on the noisy-logical integration function were much more successful overall than the associative variants. For datasets from a meta-analysis based on 17 experiments selected from 10 studies in the literature (Perales & Shanks, 2007; see also Hattori & Oaksford, 2007), the causal Bayesian models (with one or zero free parameters) performed at least as well as the most successful nonnormative model of causal learning (with four free parameters) and much better than the Rescorla-Wagner model. Thus, although both causal and associative

approaches can be given a Bayesian formulation, the empirical tests of human causal learning reported by Lu et al. favor the causal Bayesian formulation, providing further evidence for the rationality of human causal inference.

Lu et al. (2008) also evaluated structure analogs of the two causal variants of Bayesian strength models as accounts for observed structure judgments from experiments in which participants were explicitly asked to judge whether the candidate was indeed a cause. Relative to the support model, human reasoners appear to place greater emphasis on causal power and the base rate of the effect, and less emphasis on sample size.

THE INDEPENDENT CAUSAL INFLUENCE ASSUMPTION FOR DICHOTOMOUS AND CONTINUOUS OUTCOME VARIABLES

Across multiple studies on humans (Beckers, de Houwer, Pineno, & Miller, 2005; De Houwer, Beckers, & Glaudier, 2002; Lovibond et al., 2003) and even rats (Beckers, Miller, De Houwer, & Urushihara, 2006), an intriguing set of findings has emerged, showing that information regarding the additivity of the causal influences of two causes and the range of magnitudes of the outcome both influence judgments regarding unrelated candidate causes of that outcome. We illustrate the finding with parts of a broader study by Lovibond et al. (2003). In a backward blocking design, cues A and B (two food items) in combination were paired with an outcome (an allergic reaction); in a second phase, cue B alone was paired with the outcome. Thus, target cue A made no difference to the occurrence of the outcome (holding B constant, there was always an allergic reaction regardless of whether A was there). The critical manipulation in Lovibond et al. was a “pretraining compound” phase during which one group of subjects, the *ceiling* group, saw that a combination of two allergens produced an outcome at the same level (“an allergic reaction”) as a single allergen (i.e., the ceiling level). In contrast, the *nonceiling* group saw that a combination of two allergens produced a stronger reaction (“a STRONG allergic reaction”) than a single allergen (“an allergic reaction”). Following this pretraining phase, all subjects were presented with information regarding novel cues in the main training phase. Critically, the outcome in this training phase always only occurred at the intermediate level (“an allergic reaction”), both for subjects in the *ceiling* and *nonceiling* groups.

As a result of pretraining, however, subjects’ perception of the nature of the outcome in this phase would be expected to differ. For the exact same outcome, “an allergic reaction,” the only form of the outcome in that phase, whereas the *ceiling* group would perceive it to occur at the ceiling level, the *nonceiling* group would perceive it to occur at an intermediate level. As mentioned, for both groups, cue A made no difference to the occurrence of the outcome. Because the causal view represents causal relations separately from covariation, it explains why when the outcome occurs at a ceiling level, the generative effect of a cause has no observable manifestation. At a non-ceiling level, causal and associative accounts coincide: The most parsimonious explanation for no observable difference is noncausality. However, at the ceiling level, observing no difference does not allow causal inference, as explained by the causal power theory. In support of this interpretation, the mean causal rating for cue A was reliably lower for the *nonceiling* group than for the *ceiling* group.

Beckers et al. (2005) manipulated pretraining on possible levels of the outcome and on the additivity of the influences of the cues separately and found that each type of pretraining had an enormous effect on the amount of blocking. Beckers et al. (2006) obtained similar results in rats. These and other researchers have explained these results in terms of the use of propositional reasoning to draw conclusions regarding the target cue (Beckers et al., 2005, 2006; Lovibond et al., 2003; Mitchell et al., 2009). For example, a subject might reason: “If A and B are each a cause of an outcome, the outcome should occur with a greater magnitude when both A and B are present than when either occurs by itself. The outcome in fact was *not* stronger when A and B were both present as when B occurred alone. Therefore, A must not be a cause of the outcome.” These researchers have explained the impact of the pretraining in terms of learning the appropriate function for integrating the influences of multiple causes in the experimental materials (e.g., additivity vs. subadditivity) from experiences during the pretraining phase, in line with proposals by Griffiths and Tenenbaum (2005) and Lucas and Griffiths (2010).

It is important to distinguish between domain-specific integrating functions that are the outputs of causal learning, and domain-independent integrating functions that enable an output, in view of

the essential role they play in the inference process. In the causal power theory, the latter are the noisy-logicals: functions representing independent causal influence. As seen in the preceding section, whether independent causal influence is assumed in the inference process leads to different causal judgments. Moreover, independent causal influence enables compositionality. Even if we were to disregard the role of that assumption in the inference process, without it generalization of the acquired causal knowledge to new contexts would be problematic: If integrating functions were purely empirically learned, every new combination of causes, such as the combination of a target cause with unobserved causes in a new context, would require new learning (i.e., causal inference would not be compositional).

An alternative interpretation of Lovibond et al.'s (2003) results is that for all types of outcome variables, independent causal influence is always the default assumed in the causal discovery process, but the mathematical function defining independent influence differs for different types of outcome variables. For continuous outcome variables, independent causal influence is represented by additivity, as is generally known; for dichotomous outcome variables, independent causal influence is represented by the noisy-logicals, as explained earlier (see pp. 219–220 & 222–223). The unifying underlying concept is the superposition of the influences, a concept borrowed from physics. Under this interpretation, the pretraining conveys information on the nature of the outcome variable: continuous or dichotomous. Thus, subjects in their ceiling group, who received pretraining showing that two food items in combination produced an “allergic reaction” just as each item alone did, learned that the outcome is dichotomous. But subjects in their nonceiling group, who received pretraining showing that two food items in combination produced a stronger allergic reaction than each item alone, learned that the outcome is continuous.

INTERVENTION VERSUS OBSERVATION AND DIAGNOSTIC CAUSAL INFERENCE

A hallmark of a rational causal reasoner is the ability to formulate flexible and coherent answers to different causal queries. A goal of accounts of causal inference is to explain that ability. We have illustrated the causal view's answers to queries regarding causal strength and structure. (For formulations of answers to questions regarding causal attribution [how much an target outcome is due to certain causes], see Cheng & Novick, 2005; for answers to

questions regarding enabling conditions, see Cheng & Novick, 1992.) Let us consider here answers to queries involving *diagnostic* causal inference, inference from the occurrence of the effect to the occurrence of its causes. Recall Blaisdell et al.'s (2006) finding regarding rats' ability to distinguish between an event that is merely observed and one that follows an intervention. When a tone that occurred only when a light occurred during the learning phase was merely observed in the test phase, the rats in the experiment (the Observe group) nose-poked into the food bin more often than when the tone occurred immediately after that rats pressed a lever newly inserted in the test phase (the Intervene group). The Observe rats apparently diagnosed that the light must have occurred, whereas the Intervene rats diagnosed that it need not have occurred; light was never followed by food in the test phase.

Blaisdell et al.'s (2006) results were initially interpreted as support for causal Bayes nets. Note that the different diagnostic inferences in the two groups are consistent with simple deductive inference. For the Observe group, because the light was the only cause of the tone, when the tone occurred, the light must have occurred. For the Intervene group, because both the lever press and the light caused the tone, the tone occurring need not imply that the light occurred. Because causal Bayes nets (Pearl, 2000; Spirtes et al., 1993/2000) and the causal power approach (Waldmann et al., 2008) both make use of deductive inference, it is not surprising that they can also explain diagnostic reasoning.

The graphs in causal Bayes nets are assumed to satisfy the Markov condition, which states that for any variable X in the graph, conditional on its parents (i.e., the set of variables that are direct causes of X), X is independent of all variables in the graph except its descendants (i.e., its direct or indirect effects). A direct effect of X is a variable that has an arrow directly from X pointing into it, and an indirect effect of X is a variable that has a pathway of arrows originating from X pointing into it. Candidate causal networks are evaluated by assessing patterns of conditional independence and dependence entailed by the networks using the Markov and other assumptions. Candidate causal networks that are inconsistent with the observed pattern of conditional independence are eliminated, and the remaining candidate causal networks form the basis of causal judgments.

The causal Bayes nets approach explains Blaisdell et al.'s results by a distinction it makes between

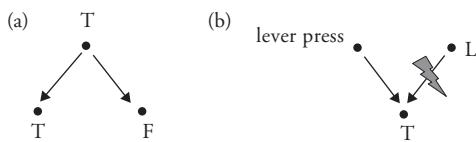


Fig. 12.5 (a) L (Light) causes T (tone) and F (food). (b) Lever press and L each causes T.

intervening to set a variable at a specific value and merely observing that value. As illustrated in Figure 12.5a, observing T allows diagnostic inference regarding L because of the arrow from L to T. But intervening to produce T severs all other incoming arrows into T, a result called graph surgery, so that the resulting causal network no longer has the arrow from L to T (see Fig. 12.5b).

Although this approach explains the results in the test phase if one assumes that the rats inferred the causal structure intended by the researchers, namely, that L is the common cause of T and F (see Fig. 12.5a), the perfect negative correlation between T and F conditional on L during the learning phase in fact violates the Markov assumption applied to this causal structure (see Rehder & Burnett, 2005; Steyvers et al., 2003 for human results indicating violations of the Markov assumption). Causal Bayes nets therefore predict from the learning phase data that there is some inhibitory connection between T and F, and that both the Intervention and Observation rats should equally avoid going to the food bin when T occurred, contrary to the responses observed.

An alternative solution, one that causal psychological theories (e.g., Cheng, 1997, 2000; Waldmann et al., 2008) inherited from traditional associative accounts (e.g., Rescorla & Wagner, 1972) is that people (and perhaps other species) incrementally construct causal networks by evaluating one (possibly conjunctive or otherwise complex) causal relation involving a single target effect at a time, while taking into consideration other causes of the effect. Motivated by consideration of limited processing capacity and of limited access to information at any one time, the incremental feature is shared by associative theorists (e.g., Jenkins & Ward, 1965 Rescorla & Wagner, 1972). Notably, whereas standard Bayes nets fail to explain Blaisdell et al.'s results, the incremental approach fully explains them. One difference is that the Markov assumption plays a different role in the latter approach: It is the consequence of the causal power assumptions

(specifically, the independence assumptions), rather than a constraint used for generating the inferences. Thus, noticing the negative correlation takes effort and thus need not occur until there is sufficient training, as is consistent with the findings in rats (Savastano & Miller, 1998; Yin, Barnet, & Miller, 1994).

In summary, the three lines of evidence just discussed all lie beyond even the augmented associative view. They converge in their support for the two leaps of faith underlying the causal view, as well as for the conviction that the causal world is logically consistent.

TIME AND CAUSALITY: MUTUAL CONSTRAINTS

We have concentrated on theoretical approaches that specify how humans take the mental leap from covariation to causation. Irrespective of any differences in theoretical perspective, all these approaches have in common that they assume that covariation can be readily assessed. This assumption is reflected in the experimental paradigms most commonly used; typically, participants are presented with evidence structured in the form of discrete, simultaneous or sequential learning trials in which each trial contains observations on whether the cause occurred and whether the effect occurred. In other words, in these tasks it is always perfectly clear whether a cause is followed by an effect on a given occasion. Such tasks grossly oversimplify the complexities of causal induction in some situations outside experimental laboratories: Some events have immediate outcomes, and others do not reveal their consequences until much later. Before an organism can evaluate whether a specific covariation licenses causal conjecture, the covariation needs to be detected and parsed in the first place.

Although the problem had been neglected for many years, the last decade has seen interesting and important developments. It has long been documented that cause-effect contiguity (one of Hume's cues toward causality) appears to be essential for causal discovery. Shanks, Pearson, and Dickinson (1989), for example, reported that in an impoverished computerized instrumental learning task, people failed to discriminate between conditions where they had strong control over an outcome ($\Delta P = .75$) and noncontingent control conditions, when their actions and the associated outcomes were separated by more than 2 seconds. In a completely different domain, Michotte (1946/1963) found that impressions of causal "launching" only

occur when the collision of the launcher with the launchee is followed immediately by motion onset in the launchee: Temporal gaps of 150 ms or more destroy the impression.

From a computational perspective, it is easy to see why delays would produce decrements in causal reasoning performance. Contiguous event pairings are less demanding on attention and memory. They are also much easier to parse. When there is a temporal delay, and there are no constraints on how the potential causes and effects are bundled, as in Shanks et al. (1989), the basic question on which contingency depends no longer has a clear answer: Should this particular instance of e be classified as occurring in the presence of c or in its absence? Each possible value of temporal lag results in a different value of contingency. The problem is analogous to that of the possible levels of abstractions of the candidate causes and the effects at which to evaluate contingency (and may have an analogous solution). Moreover, for a given e , when alternative intervening events occur, the number of hypotheses to be considered multiplies. The result is a harder, more complex inferential problem, one with a larger search space.

Buehner and May (2002, 2003, 2004) have demonstrated that prior knowledge about delayed time frames constrains the search process, such that non-contiguous relations are judged to be just as causal as contiguous ones. Buehner and McGregor (2006) have further shown that when prior assumptions about delays are sufficiently salient, contiguous relations are perceived as less causal than delayed ones—an apparent contradiction to Hume's tenets.

If causal learning operates according to the principles of Bayesian evidence integration, then these results on contiguous and delayed causation make sense: Reasoners may focus on the expected delay for a type of causal relation and evaluate observations with respect to it. In Bayesian terms, they evaluate likelihoods, the probability of the observations resulting from a hypothesis. In the earlier demonstrations of detrimental effects of delay (Michotte, 1946/1963; Shanks et al., 1989), the prior assumption would have been that there is no delay: Michotte's stimuli were simulations of well-known physical interactions (collisions), while Shanks et al. used computers, which (even in those days!) were expected to operate fast. Once these prior assumptions are modified via instructions (Buehner & May, 2002, 2003, 2004), or via constraints in the environment (Buehner & McGregor, 2006), then delayed relations pose no problem.

More recent work has found that prior expectations about time frames are relevant not only for the *extent* of delays but also with respect to their *variability*. Consider two hypothetical treatments against headache. Drug A always provides relief 2 hours after ingestion, while drug B sometimes starts working after just 1 hour, while other times it can take 3 hours to kick in. Which would we deem as a more effective drug? The answer to that question depends on how exactly temporal extent is interpreted when drawing causal conclusions. One possibility would be that causal attribution decays over time, similarly to discounting functions found in intertemporal choice (for an overview, see Green & Myerson, 2004). Under such an account, the appeal of a causal relation would decay over time according to a hyperbolic function.

One consequence of hyperbolic discounting is that variable relations may appear more attractive than stable ones, even when they are equated for mean expected delay. This conjecture is rooted in the diminishing sensitivity to delay: Variable relations accrue more net strength than constant relations matched for mean delay. And indeed, Cicerone (1976) has found that pigeons preferred variable over constant delays of reinforcement. Thus, if human causal learners approach time in a similar manner (and apply well-established principles of discounting as regards to intertemporal choice), we would expect drug B to emerge as the favorite. Interestingly, the opposite is the case: Greville and Buehner (2010) found that causal reasoners consistently prefer stable, predictable relations. Presumably we have strong *a priori* expectations that (most) causal relations are associated with a particular, relatively constant time frame. Where such expectations are violated, less learning takes place.

Cause-effect timing not only impacts assessments of causal strength but critically also constrains our ability to infer structure. Pace Hume, causes must occur before their effects, even though the intervening interval may extremely closely approximate 0 (e.g., the interval between a fist's contact with a pillow and the pillow's indentation, the interval between a cat walking into the sun and its shadow appearing on the ground). While such considerations are relatively trivial when there are only two variables involved, finding structure in multivariable causal systems gets increasingly difficult as the size of the system grows. Moreover, many structures are Markov-equivalent (Pearl, 2000), meaning that they cannot be distinguished by mere observation of the statistical patterns they produce. Lagnado

and Sloman (2004, 2006) have shown that in such situations, people rely on temporal ordering to infer causal structure. More specifically, temporal order constrains structure inference to a greater extent than the observed patterns of statistical dependencies.

As we highlighted earlier, cognitive science approaches to causality are rooted in the Humean conjecture that causality is a mental construct, inferred from hard, observable facts. Recent evidence suggests that Hume's route from sensory experience to causal knowledge is not a one-way street but in fact goes in both directions. Not only does our sensory experience determine our causal knowledge, but causal knowledge also determines our sensory experience. The latter direction of influence was first documented by Haggard, Clark, and Kalogeras (2002), who showed that sensory awareness of actions and resultant consequences are shifted in time, such that actions are perceived as later, and consequences as earlier (with reference to a baseline judgment error). Causes and effects thus mutually attract each other in our subjective experience. Originally, the effect was thought to be specific to motor action and intentional action control (Wohlschläger, Haggard, Gesierich, & Prinz, 2003). Buehner and Humphreys (2009), however, have shown that causality is the critical component of temporal binding: Intentional actions without a clear causal relation do not afford attraction to subsequent (uncaused) events. Moreover, Humphreys and Buehner (2009, 2010) have shown that the causal binding effect exists over time frames much longer than originally reported, and outside the range of motor adaptation (Stetson, Cui, Montague, & Eagleman, 2006), as would be required for action-control based approaches. Buehner and Humphreys (2010) have furthermore demonstrated a binding effect in spatial perception using Michottean stimuli—a finding that is completely outside the scope of motor-specific accounts of binding. It appears as if our perception of time and space, and our understanding of causality, mutually constrain each other to afford a maximally coherent and parsimonious experience.

Our chapter has reviewed multiple lines of evidence showing a strong preference for parsimonious causal explanations. This preference holds for scientific as well as everyday explanations. Among the many alternative representations of the world that may support predictions equally well, we select the most parsimonious. Hawking and Mlodinow (2010) note that, although people often say that Copernicus's sun-centered model of the cosmos

proved Ptolemy's earth-centered model wrong, that is not true; one can explain observations of the heavens assuming either Earth or the sun to be at rest. Likewise, although the city council of Monza, Italy, barred pet owners from keeping goldfish in curved fishbowls—on the grounds that it is cruel to give the fish a distorted view of reality through the curved sides of the bowl—the goldfish could potentially formulate a model of the motion of objects outside the bowl no less valid as ours. The laws of motion in our frame are simpler than the fish's, but theirs are potentially just as coherent and useful for prediction. But the members of the city council of Monza, like the rest of us, have such an overpowering preference for the more parsimonious model of the world that they perceive it as "truth."

Conclusions and Future Directions

Our chapter began with the question: With what cognitive assets would we endow an intelligent agent—one that has processing and informational resources similar to humans—so that the agent would be able to achieve its goals? We have taken the perspective that generalization from the learning context to the application context is central to the achievement of its goals. From this perspective, we first examined the crippling inadequacies of the associative view, which attempts to maintain objectivity by restricting its inference process to computations on observable events only. We considered variants of the associative view augmented with a special status for interventions and other principles of experimental design, in line with typical scientific causal inference.

We then considered the causal view, which resolves major apparent impasses by endowing the agent with two leaps of faith, that (1) the world is causal even though causal relations are never observable, and (2) causal laws are uniform and compositional. These empirical leaps are grounded in the conviction that existence is logically consistent. They enable the agent to incrementally construct an understanding of how the world works and coherently generalize its acquired causal knowledge. Analysis in cognitive research shows that the common belief that justifies the augmented associative view—that assumptions about independent causal influence justify the application of causal knowledge to new contexts but do not influence the output of statistical analyses—is mistaken. Likewise, the common belief that assumptions about estimations of causal strength are secondary, and do not affect judgments regarding causal structure, is mistaken.

Remarkably, observed causal judgments reveal that humans make those leaps of faith, and that their causal judgments are based on a definition of independent causal influence that is logically consistent across the learning and application contexts. The use of the sharper tool of Bayesian mathematics shows even more unequivocal support for the causal view. This tool also extends the capability to formulate answers to different kinds of causal queries.

The potential to discover how the world works must of course be accompanied by the requisite computational capabilities. We have identified three intertwined capabilities so far. The agent must be able to (1) make deductive logical inferences, (2) compute statistical regularities, and (3) represent uncertainty. The last two allow the agent to make progress in the face of errors in even its best hypotheses. The first is an essential component of the parsimony assumption and of coherent and flexible reasoning.

Three outstanding issues seem especially pertinent to us in view of our analysis and review. For each issue, a rational analysis in tandem with empirical work differentiating between alternative plausible explanations would deepen our understanding of causal learning.

1. Hypothesis revision: If causal learning is incremental, by what criteria do causal learners revise their hypotheses, and what do their criteria and revisions reveal about the intended destination of the revision process? Recent research found that for preventers with a narrow scope, which violate the independent influence assumption, people are more likely to posit a hidden cause to explain and remove the violation (Carroll & Cheng, 2010). Standard causal Bayes nets would not interpret the violation to signal a need for representation revision.
2. Category formation and causal learning: We have taken the perspective that causal discovery is the driving force underlying our mental representation of the world, not just in the sense that we need to know how things influence each other but also in the sense that causal relations define what should be considered *things* in our mental universe (Lewis, 1929). Are causal learning and category formation two aspects of the same challenge, as the goal of generalization of causal beliefs to application contexts would suggest? How do people arrive at their partitioning of the continuous stream of events into candidate causes and effects? Likewise, how do people arrive at their partitioning of events

into candidate causes and effects at particular levels of abstraction? In Lien and Cheng's (2000) experiments, human subjects were presented with causal events involving visual stimuli for which candidate-cause categories were undefined; there was no specification of either the potential critical features or the relevant level of abstraction of the features. It was found that subjects seemed to form candidate-cause categories that maximized ΔP , perhaps in an attempt to maximize the necessity and sufficiency of the cause to produce the effect in question. The topic awaits better formulations of explanations as well as additional empirical work.

3. Parsimony in causal explanations: We have encountered the critical role of parsimony in causal explanations multiple times in our chapter. Although models of parsimony (e.g., Chater & Vitányi, 2003; Lombrozo, 2007) are consistent with the psychological findings, they do not predict them. Better integration of theories of simplicity with theories and findings in causal learning would be a major advance.

Acknowledgments

Preparation of this chapter was supported by AFOSR FA 9950-08-1-0489. We thank Jessica Walker and Keith Holyoak for comments on an earlier draft.

Notes

1. More generally, for a causal tree with n nodes, the number of direct causal links would be $n - 1$ (because every node other than the root node has one and only one arrow going into it). But the number of associations between nodes (including causal ones) would be $n(n - 1)/2$, because every node in the tree is linked by an arrow to at least one other node, so that there is an non-zero association between every pair of nodes.

2. Ulrike Hahn provided this interpretation.

3. The example was provided by Sylvain Bromberger.

4. Although the theory obtains different equations for estimating generative and preventive causal powers, the choice between the two equations does not constitute a free parameter. Which of the two equations applies follows from the value of ΔP . On occasions where $\Delta P = 0$, both equations apply and make the same prediction, namely, that causal power should be 0, except in ceiling-effect situations. Here, the reasoner does have to make a pragmatic decision on whether she is evaluating the evidence to assess a preventive or generative relation, and whether the evidence at hand is meaningful for that purpose.

References

- Ahn, W.-K., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation vs. mechanism information in causal attribution. *Cognition*, 54, 299–352.

- Allan, L. G., & Jenkins, H. M. (1980). The judgment of contingency and the nature of response alternatives. *Canadian Journal of Psychology*, 34(1), 1–11.
- Anderson, J. R., & Sheu, C. F. (1995). Causal inferences as perceptual judgments. *Memory and Cognition*, 23(4), 510–524.
- Beckers, T., De Houwer, J., Pineno, O., & Miller, R. R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning. *Journal of Experimental Psychology Learning Memory and Cognition*, 31(2), 238–249.
- Beckers, T., Miller, R. R., De Houwer, J., & Urushihara, K. (2006). Reasoning rats: Forward blocking in Pavlovian animal conditioning is sensitive to constraints of causal inference. *Journal of Experimental Psychology: General*, 135, 92–102.
- Blaisdell, A. P., Sawa, K., Leising, K. J., & Waldmann, M. R. (2006). Causal reasoning in rats. *Science*, 311, 1020–1022.
- Booth, S. L., & Buehner, M. J. (2007). Asymmetries in cue competition in forward and backward blocking designs: Further evidence for causal model theory. *Quarterly Journal of Experimental Psychology*, 60, 387–399.
- Buehner, M. J., & Cheng, P. W. (1997). Causal induction: The power PC theory versus the Rescorla-Wagner model. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 55–60). Hillsdale, NJ: Erlbaum.
- Buehner, M. J., & Cheng, P. W. (2005). Causal learning. In K. J. Holyoak & R. Morrison (Eds.), *Handbook of thinking and reasoning* (pp. 143–168). Cambridge, England: Cambridge University Press.
- Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6), 1119–1140.
- Buehner, M. J., & Humphreys, G. R. (2009). Causal binding of actions to their effects. *Psychological Science*, 20(1), 1221–1228.
- Buehner, M. J., & Humphreys, G. R. (2010). Causal contraction: Spatial binding in the perception of collision events. *Psychological Science*, 21(1), 44–48.
- Buehner, M. J., & May, J. (2002). Knowledge mediates the time-frame of covariation assessment in human causal induction. *Thinking and Reasoning*, 8(4), 269–295.
- Buehner, M. J., & May, J. (2003). Rethinking temporal contiguity and the judgment of causality: Effects of prior knowledge, experience, and reinforcement procedure. *Quarterly Journal of Experimental Psychology Section A-Human Experimental Psychology*, 56A(5), 865–890.
- Buehner, M. J., & May, J. (2004). Abolishing the effect of reinforcement delay on human causal learning. *Quarterly Journal of Experimental Psychology*, 57B(2), 179–191.
- Buehner, M. J., & McGregor, S. (2006). Temporal delays can facilitate causal attribution: Towards a general timeframe bias in causal induction. *Thinking and Reasoning*, 12(4), 353–378.
- Carroll, C. D., & Cheng, P. W. (2010). The induction of hidden causes: Causal mediation and violations of independent causal influence. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 913–918). Portland, OR: Cognitive Science Society.
- Carroll, C. D., Cheng, P. W., & Lu, H. (2010). Uncertainty and dependency in causal inference. In Catrambone, R. & Ohlsson, S. (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1076–1081). Portland, OR: Cognitive Science Society.
- Chapman, G. B., & Robbins, S. J. (1990). Cue interaction in human contingency judgment. *Memory and Cognition*, 18(5), 537–545.
- Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Science*, 7, 19–22.
- Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, 104(2), 367–405.
- Cheng, P. W. (2000). Causality in the mind: Estimating contextual and conjunctive causal power. In F. Keil & R. Wilson (Eds.), *Cognition and explanation* (pp. 227–253). Cambridge, MA: MIT Press.
- Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, 99, 365–382.
- Cheng, P., & Novick, L. (2005). Constraints and nonconstraints in causal learning: Reply to White (2005) and to Luhmann and Ahn (2005). *Psychological Review*, 112(3), 694–707.
- Cicerone, R. A. (1976). Preference for mixed versus constant delay of reinforcement. *Journal of the Experimental Analysis of Behavior*, 25, 257–261.
- Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology*, 47, 109–121.
- De Houwer, J., & Beckers, T. (2002). A review of recent developments in research and theories on human contingency learning. *Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology*, 55B(4), 289–310.
- De Houwer, J., Beckers, T., & Vandorpe, S. (2002). Outcome and cue properties modulate blocking. *Quarterly Journal of Experimental Psychology*, 55A, 965–985.
- De Houwer, J., Beckers, T., & Vandorpe, S. (2005). Evidence for the role of higher-order reasoning processes in cue competition and other learning phenomena. *Learning and Behavior*, 33, 239–249.
- Dennis, M. J., & Ahn, W.-K. (2001). Primacy in causal strength judgments: The effect of initial evidence for generative versus inhibitory relationships. *Memory and Cognition*, 29(1), 152–164.
- Denniston, J. C., Savastano, H. I., & Miller, R. R. (2001). The extended comparator hypothesis: Learning by contiguity, responding by relative strength. In R. R. Mowrer & S. B. Klein (Eds.), *Handbook of contemporary learning theories* (pp. 65–117). Mahwah, NJ: Erlbaum.
- Dickinson, A., & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology*, 49B(1), 60–80.
- Garcia, J., McGowan, B. K., Ervin, F. R., & Koelling, R. A. (1968). Cues: Their relative effectiveness as a function of the reinforcer. *Science*, 160(3829), 794–795.
- Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, 118(1), 110–119.
- Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge, MA: MIT Press.
- Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111(1), 3–32.
- Green, L., & Myerson, J. (2004). A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin*, 130(5), 769–792.

- Greville, W. J., & Buehner, M. J. (2010). Temporal predictability facilitates causal learning. *Journal of Experimental Psychology: General*, 139(4), 756–771.
- Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, 51(4), 285–386.
- Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience*, 5(4), 382–385.
- Hattori, M., & Oaksford, M. (2007). Adaptive noninterventional heuristics for covariation detection in causal induction: model comparison and rational analysis. *Cognitive Science*, 31, 765–814.
- Hawking, S., & Mlodinow, L. (2010). *The grand design*. New York: Bantam Books.
- Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology*, 62, 135–163.
- Holyoak, K. J., & Hummel, J. E. (2000). The proper treatment of symbols in a connectionist architecture. In E. Dietrich & A. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines* (pp. 229–263). Mahwah, NJ: Erlbaum.
- Hume, D. (1739/1888). *A treatise of human nature*. Oxford, England: Clarendon Press.
- Humphreys, G. R., & Buehner, M. J. (2009). Magnitude estimation reveals temporal binding at super-second intervals. *Journal of Experimental Psychology: Human Perception and Performance*, 35(5), 1542–1549.
- Humphreys, G. R., & Buehner, M. J. (2010). Temporal binding of action and effect in interval reproduction. *Experimental Brain Research*, 203(2), 465–470.
- Jenkins, H., & Ward, W. (1965). Judgment of contingencies between responses and outcomes. *Psychological Monographs*, 7, 1–17.
- Karnin, L. J. (1969). Predictability, surprise, attention and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 279–296). New York: Appleton Century Crofts.
- Kant, I. (1781/1965). *Critique of pure reason*. London: Macmillan.
- Lagnado, D. A., & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(4), 856–876.
- Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology-Learning Memory and Cognition*, 32(3), 451–460.
- Lewis, C. I. (1929). *Mind and the world order*. New York: Scribner.
- Lien, Y. W., & Cheng, P. W. (2000). Distinguishing genuine from spurious causes: A coherence hypothesis. *Cognitive Psychology*, 40(2), 87–137.
- Liljeholm, M., & Cheng, P. W. (2007). When is a cause the “same?” Coherent generalization across contexts. *Psychological Science*, 18, 1014–1021.
- Lober, K., & Shanks, D. R. (2000). Is causal induction based on causal power? Critique of Cheng (1997). *Psychological Review*, 107(1), 195–212.
- Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, 55, 232–257.
- López, F. J., Cobos, P. L., & Caño, A. (2005). Associative and causal reasoning accounts of causal induction: Symmetries and asymmetries in predictive and diagnostic inferences. *Memory and Cognition*, 33, 1388–1398.
- Lovibond, P. F., Been, S. L., Mitchell, C. J., Bouton, M. E., & Frohardt, R. (2003). Forward and backward blocking of causal judgment is enhanced by additivity of effect magnitude. *Memory and Cognition*, 31(1), 133–142.
- Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, 115, 955–982.
- Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical Bayesian models. *Cognitive Science*, 34, 113–147.
- Mermin, N. D. (2005). *It's about time: Understanding Einstein's relativity*. Princeton, NJ: Princeton University Press.
- Michotte, A. E. (1946/1963). *The perception of causality* (T. R. Miles, Trans.). London: Methuen & Co.
- Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin*, 117, 363–386.
- Miller, R. R., & Matzel, L. D. (1988). The comparator hypothesis: A response rule for the expression of associations. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 22, pp. 51–92). San Diego, CA: Academic Press.
- Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, 32, 183–198.
- Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review*, 111, 455–485.
- Ochs, E., & Capps, L. (2001). *Living narrative: Creating lives in everyday storytelling*. Cambridge, MA: Harvard University Press.
- Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review*, 94(1), 61–73.
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, England: Cambridge University Press.
- Pearson, K. (1911). *The grammar of science*. (3rd ed.). New York: Meridian Books.
- Rehder, B., & Burnett, R. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, 50, 264–314.
- Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: A review and synthesis. *Psychonomic Bulletin and Review*, 14, 577–596.
- Reichenbach, H. (1956). *The direction of time*. Berkeley & Los Angeles: University of California Press.
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current theory and research* (pp. 64–99). New York: Appleton-Century Crofts.
- Salmon, W. C. (1989). Four decades of scientific explanation. In P. Kitcher & W. C. Salmon (Eds.), *Minnesota studies in the philosophy of science. Vol. 13: Scientific explanation* (pp. 3–219). Minneapolis: University of Minnesota Press.
- Savastano, H. I., & Miller, R. R. (1998). Time as content in Pavlovian conditioning. *Behavioural Processes*, 44(2), 147–162.
- Schustack, M. W., & Sternberg, R. J. (1981). Evaluation of evidence in causal inference. *Journal of Experimental Psychology: General*, 110, 101–120.
- Shanks, D. R. (1985). Forward and backward blocking in human contingency judgement. *Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology*, 37B(1), 1–21.
- Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. In G. H. Bower (Ed.), *Psychology of*

- learning and motivation—advances in research and theory* (Vol. 21, pp. 229–261). San Diego, CA: Academic Press.
- Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Temporal contiguity and the judgment of causality by human subjects. *Quarterly Journal of Experimental Psychology Section B- Comparative and Physiological Psychology*, 41(2), 139–159.
- Shepard, R. N. (2008). The step to rationality: the efficacy of thought experiments in science, ethics, and free will. *Cognitive Science*, 32, 3–35.
- Sloman, S. (2005). *Causal models: How we think about the world and its alternatives*. New York: Oxford University Press.
- Spirites, P., Glymour, C., & Scheines, R. (1993/2000). *Causation, prediction and search* (2nd ed.). Boston, MA: MIT Press.
- Stetson, C., Cui, X., Montague, P. R., & Eagleman, D. M. (2006). Motor-sensory recalibration leads to an illusory reversal of action and sensation. *Neuron*, 51(5), 651–659.
- Steyvers, M., Tenenbaum, J. B., Wagenmakers, E-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, 27, 453–489.
- Stout, S. C., & Miller, R. R. (2007). Sometimes-competing retrieval (SOCR): A formalization of the comparator hypothesis. *Psychological Review*, 114(3), 759–783.
- Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure learning in human causal induction. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural processing systems* (Vol. 13, pp. 59–65). Cambridge, MA: MIT Press.
- Tenenbaum, J. B., Kemp, C., & Griffiths, T., & Goodman, N. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331, 1279–1285.
- Thagard, P. (2000). Explaining disease: Correlations, causes, and mechanisms. In F. Keil & R. Wilson (Eds.), *Cognition and explanation* (pp. 227–253). Cambridge, MA: MIT Press.
- Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation*, 25(2), 127–151.
- Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(1), 53–76.
- Waldmann, M. R. (2001). Predictive versus diagnostic causal learning: Evidence from an overshadowing paradigm. *Psychonomic Bulletin and Review*, 8, 600–608.
- Waldmann, M. R., Cheng, P. W., Hagnay, Y., & Blaisdell, A. P. (2008). Causal learning in rats and humans: A minimal rational model. In N. Chater & M. Oaksford (Eds.), *Rational models of cognition* (pp. 453–484). Oxford, England: Oxford University Press.
- Waldmann, M. R., & Hagnay, Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(2), 216–227.
- Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, 121(2), 222–236.
- Waldmann, M. R., & Holyoak, K. J. (1997). Determining whether causal order affects cue selection in human contingency learning: Comments on Shanks and Lopez (1996). *Memory and Cognition*, 25(1), 125–134.
- White, P. A. (2002). Causal attribution from covariation information: The evidential evaluation model. *European Journal of Social Psychology*, 32(5), 667–684.
- Wohlschläger, A., Haggard, P., Gesierich, B., & Prinz, W. (2003). The perceived onset time of self- and other-generated actions. *Psychological science*, 14(6), 586–591.
- Wu, M., & Cheng, P. W. (1999). Why causation need not follow from statistical association: Boundary conditions for the evaluation of generative and preventive causal powers. *Psychological Science*, 10(2), 92–97.
- Yin, H., Barnet, R. C., & Miller, R. R. (1994). Second-order conditioning and Pavlovian conditioned inhibition: Operational similarities and differences. *Journal of Experimental Psychology: Animal Behavior Processes*, 20, 419–428.
- Yuille, A. L., & Lu, H. (2008). The noisy-logical distribution and its application to causal inference. In J. C. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems* (Vol. 20, pp. 1673–1680). Cambridge, MA: MIT Press.
- Zimmer-Hart, C. L., & Rescorla, R. A. (1974). Extinction of Pavlovian conditioned inhibition. *Journal of Comparative and Physiological Psychology*, 86, 837–845.

Analogy and Relational Reasoning

Keith J. Holyoak

Abstract

Analogy is an inductive mechanism based on structured comparisons of mental representations. It is an important special case of role-based relational reasoning, in which inferences are generated on the basis of patterns of relational roles. Analogical reasoning is a complex process involving retrieval of structured knowledge from long-term memory, representing and manipulating role-filler bindings in working memory, identifying elements that play corresponding roles, generating new inferences, and learning abstract schemas. For empirical analogies, analogical inference is guided by causal knowledge about how the source analog operates. Simpler types of relation-based transfer can be produced by relational priming. Human analogical reasoning is heavily dependent on working memory and other executive functions supported by the prefrontal cortex, with the frontopolar subregion being selectively activated when multiple relations must be integrated to solve a problem.

Key Words: analogy, role-based relational reasoning, IQ, metaphor, induction, neuroimaging, frontal cortex, symbolic connectionism, mapping, retrieval, inference, schemas, System 1, System 2, causal models, relational priming, cognitive development

Introduction

Two situations are analogous if they share a common pattern of *relationships* among their constituent elements, even though the elements themselves differ across the two situations. Identifying such a common pattern requires *comparison* of the situations. Analogy involves some of the same processes as do judgments of similarity (see Goldstone & Son, Chapter 10). Typically one analog, termed the *source* or *base*, is more familiar or better understood than the second analog, termed the *target*. By “better understood,” we mean that the reasoner has prior knowledge about functional relations *within* the source analog—beliefs that certain aspects of the source have causal, explanatory, or logical connections to other aspects (Hesse, 1966). This asymmetry in initial knowledge provides the basis for analogical transfer—using the source to generate

inferences about the target. For example, the earliest major scientific analogy, dating from the era of imperial Rome (see Holyoak & Thagard, 1995), led to a deeper understanding of sound (the target) in terms of water waves (the source). Sound is analogous to water waves in that sound exhibits a pattern of behavior corresponding to that of water waves: propagating across space with diminishing intensity, passing around small barriers, rebounding off of large barriers, and so on. The perceptual features are very different (water is wet, air is not), but the underlying pattern of relations among the elements is similar. In this example, like most analogies involving empirical phenomena, the key functional relations involve causes and their effects (see Cheng & Buehner, Chapter 12). By transferring knowledge about causal relations, the analogy provides a new explanation of why various phenomena occur (see

Lombrozo, Chapter 14). Analogy is an inductive process, and hence analogical inferences are inevitably uncertain. The wave analogy for sound proved successful; an alternative “particle” analogy did not.

In this chapter I will focus on analogy as a key example of the broader concept of *role-based relational reasoning*. After a brief review of the history of research on analogy and related concepts, such as metaphor, I will describe current views using the framework of Marr’s (1982) levels of analysis. Next, I will survey research on major subprocesses (retrieval, mapping, inference, and schema induction). This review includes both intentional and unintentional types of relational transfer, and the development of analogical abilities over the course of childhood. Finally, again applying Marr’s framework, I will consider open issues and directions for future research.

Role-Based Relational Reasoning

Analogy is a prime example of role-based relational reasoning (Penn, Holyoak, & Povinelli, 2008), as its full power depends on explicit relational representations (see Doumas & Hummel, Chapter 5). Such representations distinguish relational roles from the entities that fill those roles, while coding the bindings of entities to their specific roles. Humans are capable of making inferences about entities that cannot be reliably assigned to relational roles solely on the basis of perceptual properties. In the context of the original wave analogy, water is similar to air because each serves as a medium for the transmission of waves. The wave analogy was later extended from transmission of sound to transmission of light, and ultimately it developed into an abstract *schema*, or relational category. As another example of a relational category, something fills the role of “barrier” if it blocks the passage of something else, regardless of what type of entity the “barrier” is (perhaps a landslide, perhaps poverty). If something is known to be a barrier, its binding to that relational role is enough to infer that its removal would end the blockage. Whether any other species is capable of role-based relational reasoning is a matter of debate (see Penn & Povinelli, Chapter 27).

As the earlier examples illustrate, role-based relational reasoning is broader than reasoning by analogy between specific cases (Halford, Wilson, & Phillips, 2010). More general concepts and categories are often defined at least in part by relations (e.g., *barrier*, *parent*, *catalyst*; see Markman & Stilwell, 2001; also Rips et al., Chapter 11). Reasoning based on rules (Smith, Langston, & Nisbett, 1992), including

deductive inference (see Evans, Chapter 8; Johnson-Laird, Chapter 9), also depends critically on relations. The core property of role-based relational reasoning is that inferences about elements depend on commonalities (and sometimes differences) in the roles they play, rather than solely on perceptual features of individual elements. Although various types of inferences have this basic character, analogical inferences are especially flexible, as I will discuss in more detail later.

Functions and Processes of Analogical Reasoning

The content of analogical reasoning is extremely diverse (Holyoak & Thagard, 1995). Analogies have figured prominently in science (see Dunbar & Klahr, Chapter 35) and mathematics (Pask, 2003), and they are often used in everyday problem solving (see Bassok & Novick, Chapter 21) as well as creative cognition (Smith & Ward, Chapter 23). In legal reasoning, the use of legal precedents (relevant past cases) to help decide a new case is a special case of analogical reasoning (see Spellman & Schauer, Chapter 36). Analogies can function to sway emotions (Goode, Dahl, & Moreau, 2010; Thagard & Shelley, 2001), to influence political views (Blanchette & Dunbar, 2001; Khong, 1992), to guide consumer decisions (Markman & Loewenstein, 2010; see Lowenstein, Chapter 38), and to teach mathematics (Richland, Zur, & Holyoak, 2007). Analogy is sometimes used as part of a rational argument (Bartha, 2010; see Hahn & Oaksford, Chapter 15), using systematic connections between the source and target to generate and support plausible (though fallible) inferences about the latter.

Figure 13.1 sketches the major component processes in analogical transfer (see Carbonell, 1983; Gentner, 1983; Gick & Holyoak, 1980, 1983; Novick & Holyoak, 1991). Typically, a target situation serves as a retrieval cue for a potentially useful source analog. It is then possible to establish a *mapping*—a set of systematic correspondences that serve to align the elements of the source and target. Based on the mapping, coupled with the relevance relations within the source, it is possible to elaborate the representation of the target and derive new inferences. In the aftermath of analogical reasoning about a pair of cases, some form of relational generalization may take place yielding a more abstract schema for a category of situations (as in the case of the evolving “wave” concept), of which the source and target are both instances.

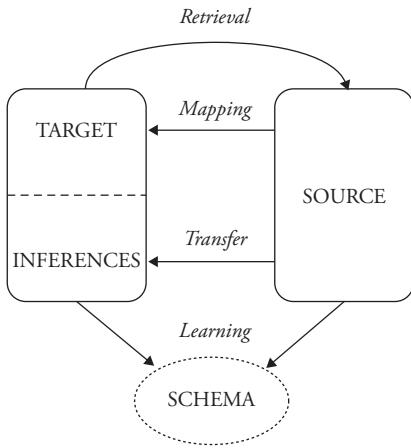


Fig. 13.1 Major components of analogical reasoning.

A Capsule History

The history of the study of analogy includes three interwoven streams of research, which respectively emphasize analogy in relation to psychometric measurement of intelligence, to metaphor and language, and to the representation of knowledge.

Psychometric Tradition

Work in the psychometric tradition focuses on four-term or “proportional” analogies, in the form A:B::C:D, such as HAND:FINGER::FOOT:?, where the problem is to infer the missing D term (TOE) that is related to C in the same way B is related to A. The pair A:B thus plays the role of source analog, and C:D that of target. Proportional analogies were discussed by Aristotle (see Hesse, 1966), and in the early decades of modern psychology became a centerpiece of efforts to define and measure intelligence. Charles Spearman (1923, 1927) argued that the best account of observed individual differences in cognitive performance was based on a general or *g* factor, with the remaining variance being unique to the particular task. He reviewed several studies that revealed high correlations between performance in solving analogy problems and the *g* factor. Spearman’s student John C. Raven (1938) developed the Raven’s Progressive Matrices Test (RPM), which requires selection of a geometric figure to fill an empty cell in a two-dimensional matrix (typically 3 x 3) of such figures. Much like a geometric proportional analogy, the RPM requires participants to extract and apply information based on visuospatial relations. (See Hunt, 1974, and Carpenter, Just, & Shell, 1990, for analyses of strategies for solving RPM problems.) The RPM proved to be an especially pure measure of *g*.

Raymond Cattell (1971), another student of Spearman, elaborated his mentor’s theory by distinguishing between two components of *g*: *crystallized* intelligence, which depends on previously learned information or skills, and *fluid* intelligence, which involves reasoning with novel information. As a form of inductive reasoning, analogy would be expected to require fluid intelligence. Cattell confirmed Spearman’s (1946) observation that analogy tests and the RPM provide sensitive measures of *g*, clarifying that they primarily measure fluid intelligence (although verbal analogies based on difficult vocabulary items also depend on crystallized intelligence). Figure 13.2 graphically depicts the centrality of RPM performance in a space defined by individual differences in performance on various cognitive tasks. Note that numerical, verbal, and geometric analogies cluster around the RPM at the center of the figure.

Because four-term analogies and the RPM are based on small numbers of relatively well-specified elements and relations, it is possible to systematically manipulate the complexity of such problems and analyze performance (based on response latencies and error rates) in terms of component processes (e.g., Mulholland, Pellegrino, & Glaser, 1980; Sternberg, 1977). The earliest computational models of analogy were developed for four-term analogy problems (Evans, 1968; Reitman, 1965). The basic components of these models were elaborations of those proposed by Spearman (1923), including encoding of the terms, accessing a relation between the A and B terms, and evoking a comparable relation between the C and D terms. As we will discuss later, four-term analogies and the RPM have proved extremely useful in recent work on the cognitive neuroscience of analogy.

Metaphor

Analogy is closely related to metaphor and related forms of symbolic expression that arise in everyday language (e.g., “the evening of life,” “the idea blossomed”), in literature (Holyoak, 1982), the arts, and cultural practices such as ceremonies (see Holyoak & Thagard, 1995, ch. 9). Like analogy in general, metaphors are characterized by an asymmetry between target (conventionally termed “tenor”) and source (“vehicle”) domains (e.g., the target/tenor in “the evening of life” is life, which is understood in terms of the source/vehicle of time of day). In addition, a mapping (the “grounds” for the metaphor) connects the source and target,

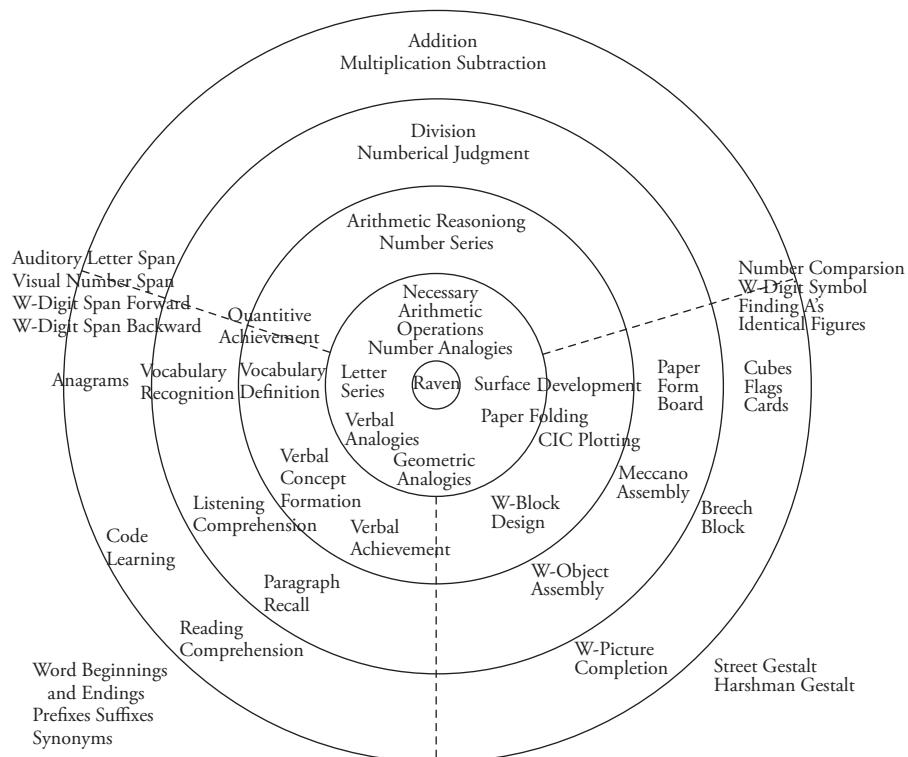


Fig. 13.2 Multidimensional scaling solution based on intercorrelations among the Raven's Progressive Matrices test, analogy tests, and other common tests of cognitive function. (Reprinted by permission from Snow, Kyllonen, & Marshalek, 1984, p. 92.)

allowing the domains to interact to generate a new conceptualization (Black, 1962). Metaphors are a special kind of analogy, in that the source and target domains are always semantically distant (Gentner, 1982; Gentner, Falkenhainer, & Skorstad, 1988), and the two domains are often blended rather than simply mapped (e.g., in “the idea blossomed,” the target is directly described in terms of an action term derived from the source). In addition, metaphors are often combined with other symbolic “figures,” especially metonymy (substitution of an associated concept). For example, “sword” is a metonymic expression for weaponry, derived from its ancient association as the prototypical weapon; “Raising interest rates is the Federal Reserve Board’s sword in the battle against inflation” extends the metonymy into metaphor.

Fauconnier and Turner (1998; Fauconnier, 2001) have analyzed complex conceptual blends that are akin to metaphor. A typical example is a description of the voyage of a modern catamaran sailing from San Francisco to Boston, which was attempting to beat the speed record set by a clipper ship that had sailed the same route over a century earlier. A magazine account written during the catamaran’s voyage

said the modern boat was “barely maintaining a 4.5-day lead over the ghost of the clipper *Northern Light*.” Fauconnier and Turner observed that the magazine writer was describing a “boat race” that never took place in any direct sense; rather, the writer was blending the separate voyages of the two ships into an imaginary race. The fact that such conceptual blends are so natural and easy to understand attests to the fact that people can readily comprehend novel metaphors.

Lakoff and Johnson (1980; also Lakoff & Turner, 1989) have argued that much of human experience, especially its abstract aspects, is grasped in terms of broad conceptual metaphors (e.g., events occurring in time are understood by analogy to objects moving in space). Time, for example, is understood in terms of objects in motion through space, as in expressions such as “My birthday is fast approaching” and “The time for action has arrived.” (See Boroditsky, 2000, for evidence of how temporal metaphors influence cognitive judgments.) As Lakoff and Turner (1989) pointed out, the course of a life is understood in terms of time in the solar year (youth is springtime, old age is winter). Life is also conventionally conceptualized as a journey. Such conventional metaphors can still

be used in creative ways, as illustrated by Robert Frost's famous poem, "The Road Not Taken":

Two roads diverged in a wood, and I—
I took the one less traveled by,
And that has made all the difference.

According to Lakoff and Turner, comprehension of this passage depends on our implicit knowledge of the metaphor that life is a journey. This knowledge includes understanding several interrelated correspondences (e.g., a person is a traveler, purposes are destinations, actions are routes, difficulties in life are impediments to travel, counselors are guides, and progress is the distance traveled).

Psychological research has focused on demonstrations that metaphors are integral to everyday language understanding (Glucksberg, Gildea, & Bookin, 1982; Keysar, 1989). There has been a debate about whether metaphor is better conceptualized as a kind of analogy (Wolff & Gentner, 2000) or a kind of categorization (Glucksberg & Keysar, 1990; Glucksberg, McClone, & Manfredi, 1997). A likely resolution is that novel metaphors are interpreted by much the same processes as are analogies, whereas more conventional metaphors are interpreted as more general schemas (Gentner & Bowdle, 2008; see discussion of schemas later in this chapter).

Knowledge Representation

The most important influence on analogy research in the cognitive-science tradition has been concerned with the representation of knowledge within computational systems (see Markman, Chapter 4). Many seminal ideas were developed by the philosopher Mary Hesse (1966), who was in turn influenced by Aristotle's discussions of analogy in scientific classification and Black's (1962) interactionist view of metaphor. Hesse placed great stress on the *purpose* of analogy as a tool for scientific discovery and conceptual change, and on the close connections between causal relations and analogical mapping. In the 1970s, work in artificial intelligence and psychology focused on the representation of complex knowledge of the sort used in scientific reasoning, problem solving, story comprehension, and other tasks that require structured knowledge. A key aspect of structured knowledge is that elements can be flexibly *bound* into the roles of relations. For example, "dog bit man" and "man bit dog" have the same elements and the same relation, but the role bindings have been reversed, radically altering the overall

meaning. How the mind and brain accomplish role binding is thus a central problem to be solved by any psychological theory that involves structured knowledge, including any theory of analogy (see Doumas & Hummel, Chapter 5).

In the 1980s, a number of cognitive scientists recognized the centrality of analogy as a tool for discovery, as well as its close connection with theories of knowledge representation. Winston (1980), guided by Minsky's (1975) treatment of knowledge representation, built a computer model of analogy that highlighted the importance of causal relations in guiding analogical inference. Other researchers in artificial intelligence also began to consider the use of complex analogies in reasoning and learning (Kolodner, 1983; Schank, 1982), leading to an approach to artificial intelligence termed *case-based reasoning* (Kolodner, 1993).

Meanwhile, cognitive psychologists began to consider analogy in relation to knowledge representation and eventually to integrate computational modeling with detailed experimental studies of human analogical reasoning. Gentner (1982, 1983; Gentner & Gentner, 1983) investigated the role of analogy in understanding scientific topics. She emphasized that in analogy, the key similarities involve relations that hold within the domains (e.g., the flow of electrons in an electrical circuit is analogically similar to the flow of people in a crowded subway tunnel), rather than in features of individual objects (e.g., electrons do not resemble people). Moreover, analogical similarities often depend on *higher order* relations—relations *between* relations. For example, adding a resistor to a circuit *causes* a decrease in flow of electricity, just as adding a narrow gate in the subway tunnel would decrease the rate at which people pass through (where *causes* is a higher order relation). In her structure-mapping theory, Gentner proposed that analogy entails finding a structural alignment, or mapping, between elements of the two domains. In this theory, a "good" alignment between two representational structures is characterized by a high degree of structural parallelism (consistent, one-to-one correspondences between mapped elements) and of systematicity—an implicit preference for deep, interconnected systems of relations governed by higher order relations, such as causal, mathematical, or other functional relations.

Holyoak and his colleagues (1985; Gick & Holyoak, 1980, 1983; Holyoak & Koh, 1987) focused on the role of analogy in problem solving, with a strong concern for the role of pragmatics in

analogy—how causal relations that impact current goals and context guide the interpretation of an analogy. Holyoak and Thagard (1989, 1995) developed an approach to analogy in which several factors were viewed as jointly constraining analogical reasoning. According to their *multiconstraint* theory, people implicitly favor mappings that maximize structural parallelism (in agreement with Gentner's, 1983, structure-mapping theory), but that also maximize direct similarity of corresponding elements and relations, and that give priority to pragmatically important elements (i.e., those functionally related to achieving a goal). The theory further specified how the joint influence of these constraints, which often converge but sometimes conflict, might be adjudicated by a process of constraint satisfaction.

Other early work dealt with role-based relational reasoning more broadly. Gick and Holyoak (1983) provided evidence that analogical comparisons can provide the seed for forming new relational categories, by abstracting the relational correspondences between examples to form a schema for a class of problems. Halford (1993; Halford & Wilson, 1980) argued that the development of the ability to map relational structures is central to cognitive development. More generally, role-based relational reasoning came to be viewed as a central part of human induction (Holland, Holyoak, Nisbett, & Thagard, 1986; see Markman, Chapter 4; Doumas & Hummel, Chapter 5), with close ties to other basic thinking processes, including causal inference (Cheng & Buehner, Chapter 12), categorization (Rips et al., Chapter 11), and problem solving (Bassok & Novick, Chapter 21).

Relational Reasoning: Levels of Analysis

To provide an overview of current conceptions of analogy, I will focus on three questions: What are the functions of human relational reasoning, by what algorithms is it achieved, and how is it implemented in the brain? These questions instantiate Marr's (1982) three levels of analysis: computation, representation and algorithm, and implementation. Cognitive scientists and cognitive neuroscientists have addressed all three levels, focusing on analogical reasoning as a central example.

Computational Goal

At the computational level, analogies are used to achieve the goals of the reasoner (Holyoak, 1985). These goals are diverse—forming and evaluating hypotheses, solving problems, understanding new

concepts, winning arguments, and so on. The focus here will be on the role of analogies and relational reasoning in pursuing what Molden and Higgins (Chapter 20) term a basic “nondirectional outcome goal”: truth, coupled with relevance to current goals. The scientist seeking a good theory, the architect creating a building design that meets the client's needs, the child trying to understand how the world works, are all basically motivated to use their prior knowledge—including specific analogs—to make true and useful inferences. Of course, inductive uncertainty is inevitable, and analogies can potentially mislead. Nonetheless, the rational reasoner (see Stanovich, Chapter 22) will use analogies to reach rationally justified inferences relevant to achieving current goals—inferences that though fallible are at least plausible, and are accompanied by an appropriate sense of their degree of uncertainty (Bartha, 2010; Lee & Holyoak, 2008).

The overarching goal of making true and useful inferences underlies the major constraints on analogical inference that have been discussed in the literature. For an analogy to be successful, the structure and content of the source must provide a good model to use in elaborating the representation of the target. To the extent that the reasoner understands the functional structure of the source (i.e., what aspects depend on which other aspects), it will be possible to focus on goal-relevant information in it while “backgrounding” other details. The functional structure may take different forms. For example, in a mathematical analogy, the functional structure will involve the mathematical or logical properties that justify a conclusion (see Bartha, 2010). For empirical knowledge (i.e., knowledge about how things happen in the real world), the aim is to transfer a *causal model* (see Cheng & Buehner, Chapter 12) from source to target. In such cases the backbone of the functional structure will be cause-effect relations—“what makes what happen” in the source domain drives potential inferences about “what makes what happen” in the target.

From the perspective of Holyoak and Thagard's (1989) multiconstraint theory, the centrality of functional structure is a basic pragmatic constraint. Causal relations constitute the prime example of “higher order” relations involved in Gentner's systematicity constraint, which has been supported by experimental evidence that analogical transfer is more robust when the source includes causal structure than when it does not (Gentner & Toupin, 1986). In general, a highly systematic source will

be rich in functional structure. A high degree of structural parallelism (that is, a consistent mapping between relevant elements of the source and target) is a logical requirement if the structure of the source is to provide an appropriate model for the structure of the target. Ambiguity will be minimized if the mapping is both consistent and one to one.

Holyoak and Thagard's (1989) constraint of semantic similarity—a preference for mappings in which similar objects are placed into correspondence—also follows from the overarching goal of seeking true and goal-relevant inferences. Direct semantic similarity of elements has often been termed “surface” similarity, in contrast to the “structural” variety. In fact, this contrast has been defined in (at least) two distinct ways in the analogy literature, indicating resemblances based either (1) on features versus relations (Gentner, 1983) or (2) on functionally irrelevant versus relevant elements of the analogs (Holyoak, 1985). In general, functional structure (the latter sense) will involve not only relations (i.e., predicates that take at least two arguments; see Doumas & Hummel, Chapter 5) but also those additional elements that participate in functional relations. For example, because an orange is round, fairly small, and firm (properties usually considered to be perceptual features, not relations), it could be considered analogous to a ball for purposes of playing catch. In this example, as in most simple empirical analogies, various perceptual properties participate in relevant causal relations and hence count as “structural” by the functional definition. In general, objects that share direct similarities are likely to have similar causal properties (see Rips et al., Chapter 11). Thus, while “distant” analogies between remote domains of knowledge may be especially creative (see Smith & Ward, Chapter 23), “close” analogies in which similar entities fill corresponding roles typically provide stronger support for plausible inferences (Medin & Ross, 1989; see also Koedinger & Roll, Chapter 40).

Representation and Algorithm

Role-based relational reasoning depends on the capacity to represent structured relations in terms of their roles, to represent the bindings of entities to roles, to find systematic correspondences between a source and target based on relational structure, and to use this structure to create new propositions about the target. Because of its dependence on explicit relations, all major computational models of analogical reasoning (e.g., Falkenhainer, Forbus, &

Gentner, 1989; Halford, Wilson, & Phillips, 1998; Hofstadter & Mitchell, 1994; Holyoak & Thagard, 1989; Hummel & Holyoak, 1997, 2003; Keane & Brayshaw, 1988; Kokinov & Petrov, 2001) are based on some form of propositional representation capable of expressing role-filler bindings (see Markman, Chapter 4; Doumas & Hummel, Chapter 5).

Most algorithmic models of analogy, formalized as computer simulations, are based on traditional symbolic representations. Traditional connectionist systems, which lack the capacity to code variable bindings, have not been successful in modeling human-like relational reasoning (see Doumas & Hummel, Chapter 5), although they may well be applicable to simpler types of relational processing (see later discussion of relational priming), some of which appear to be within the capabilities of non-human animals (see Penn & Povinelli, Chapter 27). The most promising algorithmic approach to modeling human analogical reasoning in a way that makes contact with data on its neural basis is *symbolic connectionism* (Halford et al., 1998, 2010; Hummel & Holyoak, 1997, 2003; see Doumas & Hummel, Chapter 5). Models of this type aim to represent structured relations (hence “symbolic”) within relatively complex neural networks (hence “connectionist”), and furthermore aim to operate within a human-like limited-capacity working memory.

The central role of working memory and related executive processes in analogical reasoning has long been supported by research in the psychometric tradition, as described earlier (see Fig. 13.2). More recent experimental work, both with normal and brain-damaged populations, has provided further evidence. For example, Waltz, Lau, Grewal, and Holyoak (2000) asked college students to map objects in a pair of pictorial scenes while simultaneously performing a secondary task designed to tax working memory (e.g., generating random digits). Adding a dual task diminished relational responses and increased similarity-based responses. A manipulation that increases people's anxiety level (performing mathematical calculations under speed pressure prior to the mapping task) yielded a similar shift in mapping responses (Tohill & Holyoak, 2000; also Feldman & Kokinov, 2009). Most dramatically, degeneration of the frontal lobes radically impairs relation-based mapping (Morrison et al., 2004; Waltz et al., 1999). These and many other findings (e.g., Cho, Holyoak, & Cannon, 2007) demonstrate that mapping on the basis of relations requires adequate working memory and

attentional resources to represent and manipulate role bindings. The neural substrate of working memory is relatively well understood (see Morrison & Knowlton, Chapter 6), and its genetic mechanisms are being actively investigated (see Green & Dunbar, Chapter 7). By connecting to working memory, symbolic-connectionist models provide a potential algorithmic “bridge” between the computational and implementational levels of analysis for role-based relational reasoning.

One example of a symbolic-connectionist model of analogy is LISA (Learning and Inference with Schemas and Analogies; Hummel & Holyoak, 1997, 2003; Doumas, Hummel, & Sandhofer, 2008). LISA (described more fully by Doumas & Hummel, Chapter 5) is based on the principles of Holyoak and Thagard’s (1989) multiconstraint theory of analogy. The model aims to provide a unified account of all the major components of analogical reasoning. LISA represents propositions using a hierarchy of distributed and localist units (see Fig. 13.5b; also Fig. 5.5 in Doumas & Hummel, Chapter 5). LISA includes both a long-term memory for propositions and concept meanings and a limited-capacity working memory. LISA’s working memory representation, which uses neural synchrony to encode role-filler bindings, provides a natural account of the capacity limits of working memory because it is only possible to have a finite number of bindings simultaneously active and mutually *out of synchrony*.

Analog retrieval is accomplished as a form of guided pattern matching. Propositions in a *driver* analog (typically the target) generate synchronized patterns of activation on the semantic units, which in turn activate propositions in *recipients*—potential source analogs residing in long-term memory. The resulting coactivity of elements of the target and a selected source, augmented with a capacity to learn which structures in the target were coactive with which in the source, serves as the basis for analogical mapping. LISA includes a set of *mapping connections* between units of the same type (e.g., object, predicate) in separate analogs. These connections grow whenever the corresponding units are active simultaneously, and thereby permit LISA to learn correspondences between structures in separate analogs. Augmented with an algorithm for self-supervised learning, the model can generate analogical inferences based on the mapping (by using the source as the driver to generate new relational structure in the target); and further augmented with an algorithm

for intersection discovery, the model provides a basis for schema induction.

LISA has been used to simulate a wide range of behavioral data on analogical reasoning in normal adults (Hummel & Holyoak, 1997, 2003). To take just one example, LISA predicts that mapping complex situations must be performed sequentially, because only a small number of propositions (two to three) can be active together in the driver analog. Mappings will be established incrementally, with early mappings constraining later ones (cf. Keane, 1997). Moreover, the success of the mapping process will depend on *which* propositions are activated together in the driver. In general, coherent, interconnected propositions (e.g., facts that are causally related) will provide more information that can be used to disambiguate a complex mapping. LISA therefore predicts that mapping will be more successful when the better-understood source acts as the driver while the less-understood target serves as recipient. Kubose, Holyoak, and Hummel (2002) showed that coherence of the driver impacts mapping, as LISA predicts. For example, people are more accurate in mapping the solved “general” story to the unsolved “tumor” problem than the reverse (see later discussion of “convergence” problems; Gick & Holyoak, 1980).

In addition to explaining phenomena concerning analogical reasoning by normal adults, LISA can account for numerous findings involving similarity judgments (Taylor & Hummel, 2009), developmental patterns (Doumas et al., 2008; Morrison, Doumas, & Richland, 2011), evidence of deficits in relational reasoning in older normal adults (Viskontas et al., 2004), and evidence of much more pronounced deficits in patients with lesions to their frontal or temporal cortex (Morrison et al., 2004).

Neural Substrate of Relational Reasoning

The implementation of role-based relational reasoning in the human brain involves a broad interconnected network of brain regions (see Morrison & Knowlton, Chapter 6). Although it is generally simplistic to identify cognitive functions with specific brain regions, several regions play major roles. The prefrontal cortex (PFC) is of central importance. The basic processes of the LISA model are closely related to known functions of PFC, in particular rapid learning (e.g., Asaad et al., 1998; Cromer, Machon, & Miller, 2011) and inhibitory control (e.g., Miller & Cohen, 2001). Other important brain regions include the hippocampus (critical

for storage and retrieval of episodic knowledge), the anterior temporal cortex (storage of semantic information, including semantic relations), and the parietal cortex (representation of spatial relations).

Role of Prefrontal Cortex in Relational Integration and Interference Control

The prefrontal cortex plays a central role in relational reasoning. It has been argued that this area underlies the fluid component of Spearman's g factor in intelligence (Duncan et al., 2000), and it supports the executive functions of working memory and cognitive control. With respect to relational reasoning, the PFC is critical in the maintenance and active manipulation of relations and role bindings (Knowlton & Holyoak, 2009; Robin & Holyoak, 1995). Waltz et al. (1999) found that patients with frontal-lobe damage showed a marked deficit in solving problems of the Raven's-matrix type problems that required integration of two relations compared to normal controls and patients with anterior temporal lobe damage. The frontal-lobe patients performed comparably to the other groups on less complex problems that could be solved using zero or one relation. These findings imply that prefrontal cortex is critical for the integration of multiple relations.

Other neuropsychological studies have examined the role of the PFC in controlling interference from distracting information during analogical reasoning. Morrison et al. (2004) tested patients with either frontal or temporal damage, as well as age-matched controls, on a verbal analogy task. Four-term analogy problems of the form A:B::C:D or D' were employed, where D is the analogical answer and D' is a nonanalogical foil. A semantic facilitation index (SFI) was calculated for each problem to characterize the association of the correct relational pair (C:D) relative to the distractor pair (C:D'). For example, for the problem *play:game::give: (party or take)*, the C:D pair (*give:party*, the correct analogical answer) is less associated than is the C:D' pair (*give:take*, the nonanalogical foil), yielding a negative SFI for the problem. The problems were divided into those with negative SFI, neutral SFI, and positive SFI in order to examine the effect of semantic interference on the ability to identify the analogical answer. Frontal patients were selectively impaired in the negative SFI condition relative to the positive and neutral SFI conditions, consistent with the hypothesis that the frontal cortex is necessary for control of interference. In contrast, temporal patients showed a more uniform decline in verbal analogy performance

across all three conditions, due to their loss of the conceptual information necessary to encode the relations in the analogy problem. Using four-term picture analogies, Krawczyk et al. (2008) also found that frontal patients are especially impaired on problems that include semantically related distractors.

Functional Decomposition of Prefrontal Cortex in Reasoning Tasks

Several neuroimaging studies using functional magnetic resonance imaging (fMRI) have manipulated relational complexity using variants of Raven's Progressive Matrices (RPM) problems, similar to those used by Waltz et al. (1999) in their neuropsychological studies. For matrix problems, relational integration has been shown to consistently activate prefrontal regions. In particular, bilateral middle and inferior frontal gyri, as well as parietal and occipital regions, have been found to increase activity when multiple relations must be integrated in order to arrive at a solution, compared to problems that require processing of only a single relation (Christoff et al., 2001; Kroger et al., 2002).

Among these regions, which constitute a network commonly activated in visuospatial working memory tasks, the activation pattern of the most anterior part of the PFC (termed *frontopolar*, or *rostrolateral*) has been particularly noteworthy. Christoff et al. (2001) found that the left frontopolar region remained preferentially activated even after controlling for the influence of increased problem-solving time (also Kroger et al., 2002). Similarly, studies of verbal analogical reasoning have distinguished neural substrates of reasoning from semantic processing demands within working memory. Activation in the left frontopolar region increases selectively when making judgments of analogical similarity compared to processing of semantic associations or categories (Bunge et al., 2005; Green et al., 2006; Wendelken et al., 2008). Moreover, frontopolar activation selectively increases when the semantic distance between the A:B and C:D pairs in a "true/false" verbal analogy problem is increased (Green et al., 2010), a manipulation that may increase the demands on relational processing.

Thus, based on a substantial body of findings involving solution of different types of relational reasoning problems, the frontopolar region appears to play a special role in the process of integrating multiple relational representations to arrive at a solution. Other subregions of PFC subserve additional processes involved in relational reasoning. Cho et al. (2010)

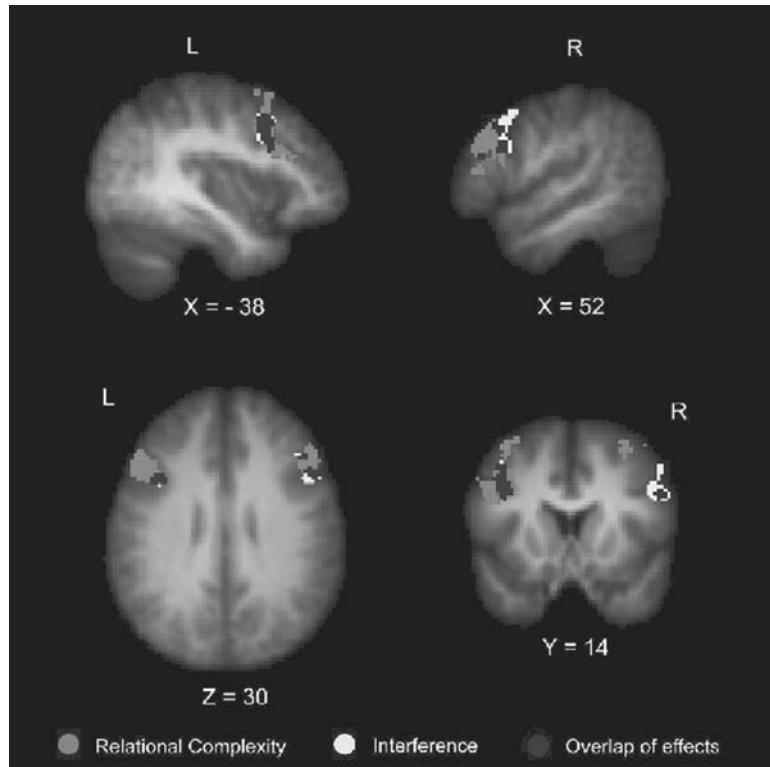


Fig. 13.3 Neuroimaging results from Cho et al. (2010). Regions showing the main effects of relational complexity (shown in red), interference (shown in yellow; small volume corrected, uncorrected cluster-forming threshold $T > 2.3$, corrected cluster extent significance threshold, $p < .05$), and regions where main effects overlapped (blue) within an a priori defined anatomical ROI mask of the bilateral MFG and IFG pars opercularis and pars triangularis. R, right; L, left. Coordinates are in MNI space (mm). (Reprinted by permission.) See color figure.

performed an fMRI study using four-term analogy problems based on cartoon figures and observed a partial dissociation between cortical regions sensitive to increase in demands on integration of multiple goal-relevant relations versus control of interference from goal-irrelevant relations (see Fig. 13.3). Problems requiring greater interference control selectively activated portions of the inferior frontal gyrus.

Component Processes of Analogical Reasoning

Let us now consider the subprocesses of relational reasoning in greater detail. A canonical instance of analogical reasoning involves (1) using retrieval cues provided by the target situation to access one or more source analogs in memory, (2) finding a mapping between a source and target, (3) using the mapping together with the functional structure of the source to make inferences about the target, and (4) generalizing the functional structure of the analogs (see Fig. 13.1).

A Paradigm for Studying Analogical Transfer

To study the entire process of analogical reasoning, including the retrieval of a source analog from

long-term memory, a key requirement is to ensure that one or more source analogs are in fact potentially available to the reasoner. Gick and Holyoak (1980, 1983) introduced a general laboratory paradigm for investigating analogical transfer in the context of problem solving. The basic procedure was to first provide people with a source analog in the guise of some incidental context, such as an experiment on “story memory.” Later, participants were asked to solve a problem that was in fact analogous to the story they had studied earlier. The questions of central interest were (1) whether people would spontaneously notice the relevance of the source analog and use it to solve the target problem, and (2) whether they could solve the analogy once they were cued to consider the source. Spontaneous transfer of the analogous solution implies successful retrieval and mapping; cued transfer implies successful mapping once the need to retrieve the source has been removed.

The source analog used by Gick and Holyoak (1980) was a story about a general who is trying to capture a fortress controlled by a dictator and needs to get his army to the fortress at full strength. Since the entire army could not pass safely along any single road, the general sends his men in small

groups down several roads simultaneously. Arriving at the same time, the groups join up and capture the fortress.

A few minutes after reading this story under instructions to read and remember it (along with two other irrelevant stories), participants were asked to solve a tumor problem (Duncker, 1945), in which a doctor has to figure out how to use rays to destroy a stomach tumor without injuring the patient in the process. The crux of the problem is that it seems that the rays will have the same effect on the healthy tissue as on the tumor—high intensity will destroy both, low intensity neither. The key issue is to figure out how the rays can be made to selectively impact the tumor while sparing the surrounding tissue. The source analog, if it can be retrieved and mapped, can be used to generate a “convergence” solution to the tumor problem, one that parallels the general’s military strategy: Instead of using a single high-intensity ray, the doctor could administer several low-intensity rays at once from different directions. In that way each ray would be at low intensity along its path, and hence harmless to the healthy tissue, but the effects of the rays would sum to achieve the effect of a high-intensity ray at their focal point, the site of the tumor.

When Gick and Holyoak (1980) asked college students to solve the tumor problem, without a source analog, only about 10% of them produced the convergence solution. When the general story had been studied, but no hint to use it was given, only about 20% of participants produced the convergence solution. In contrast, when the same participants were then given a simple hint that “you may find one of the stories you read earlier to be helpful in solving the problem,” about 75% succeeded in generating the analogous convergence solution. In other words, people often fail to notice superficially dissimilar source analogs that they could readily use. On occasions when a person did not notice the relevance of the remote source analog, he or she sometimes reported a feeling of insight (see van Steenburgh et al., Chapter 24).

Accessing Analogs in Long-Term Memory

This gap between the difficulty of retrieving remote analogs and the relative ease of mapping them has been replicated many times, both with adults (Gentner, Rattermann, & Forbus, 1993; Holyoak & Koh, 1987; Ross, 1987, 1989; Spencer & Weisberg, 1986) and with young children (Chen, 1996; Holyoak, Junn, & Billman, 1984; Tunteler &

Resing, 2002). When analogs must be cued from long-term memory, cases from a domain similar to that of the cue are retrieved much more readily than cases from remote domains (Keane, 1987; Seifert, McKoon, Abelson, & Ratcliff, 1986). For example, Keane (1987) measured retrieval of a convergence analog to the tumor problem when the source analog was studied 1–3 days prior to presentation of the target radiation problem. Keane found that 88% of participants retrieved a source analog from the same domain (a story about a surgeon treating a brain tumor), whereas only 12% retrieved a source from a remote domain (the general story). This difference in ease of access was dissociable from the ease of postaccess mapping and transfer, as the frequency of generating the convergence solution to the radiation problem once the source analog was cued was high and equal (about 86%) regardless of whether the source analog was from the same or a different domain.

The “retrieval gap” found in experimental studies of analogy is consistent with the obvious differences in the computational requirements of retrieval versus mapping. When attention is focused on two analogs, their representations will be held in working memory and explicit comparison processes can operate on relations. By definition, the question of which situations ought to be compared has already been answered. In contrast, retrieval is wide open—anything in long-term memory might potentially be relevant. As noted earlier, direct similarity of objects in the source and target is often a valid predictor of the inferential usefulness of the source. Additionally, focusing on relations in the target as retrieval cues places greater demands on working memory (see Morrison & Knowlton, Chapter 6).

Nonetheless, there is strong evidence that relational structure does play an important role in guiding analogical retrieval, both in the context of problem solving (Holyoak & Koh, 1987; Ross, 1987, 1989) and story reminding (Wharton et al., 1994; Wharton, Holyoak, & Lange, 1996). The influence of relational correspondences is greater when some degree of direct similarity of objects is also present, and when the relational correspondences favor one potential source over a nonrelational candidate competing for retrieval from long-term memory (Wharton et al., 1994). Interestingly, retrieval of verbal analogs (in the form of proverbs) is more successful when the analogs are presented in spoken rather than written form (Markman, Taylor, & Gentner, 2007), perhaps

because listening imposes a reduced processing load relative to reading. In addition, domain experts (who are more likely to focus on relevant relations as retrieval cues) are more likely than novices to access remote source analogs based on relational correspondences (Novick, 1988; Novick & Holyoak, 1991). Other evidence indicates that having people generate example cases, as opposed to simply asking them to remember cases presented earlier, can enhance structure-based access to source analogs (Blanchette & Dunbar, 2000).

UNINTENDED MEMORY ACTIVATION AND RELATIONAL PRIMING

In most of the experiments discussed so far, participants were explicitly asked to remember analogous situations stored in memory when cued with an analog, and hence were clearly aware when retrieval took place. In others (e.g., Gick & Holyoak, 1980) retrieval was not explicitly requested, but participants generally seemed to be aware of using a source analog to solve the target problem (when they in fact did so). Under some circumstances, however, people may use relations as cues to access information in long-term memory even when they have not been asked to do so. Moreover, in some cases they may not be aware that a previously encountered analog is guiding their current processing of a new example. Schunn and Dunbar (1996) performed a study in which during an initial session involving a problem in biochemistry, some subjects learned that addition of an inhibitory enzyme decreased virus reproduction. In a subsequent session the following day these same subjects were asked to solve a molecular-genetics problem, which involved an analogous inhibitory gene. Schunn and Dunbar found that subjects who had been exposed to the concept of inhibition in the initial session were more likely than control subjects to develop a solution based on inhibition for the transfer problem, even though experimental subjects evinced no signs of awareness that the earlier virus problem had influenced their solution to the gene problem. Similarly, Day and Goldstone (2011) found that participants were able to transfer strategies learned from a perceptually concrete simulation of a physical system to a task with very dissimilar content and appearance. Although recognition of the analogy between the tasks was associated with better overall performance, transfer (i.e., application of an analogous strategy) was not related to such recognition (see also Day & Gentner, 2007; Wharton & Lange, 1994).

Such apparently unintended transfer likely involves a different mechanism than does deliberate analogical mapping and inference. As we will see later, intentional relational transfer makes heavy demands on working memory and appears to be a paradigmatic example of what is sometimes termed *explicit* or System 2 processing (see Evans, Chapter 8). But as Schunn and Dunbar (1996) argued, some forms of relational transfer may be more akin to *priming*, typically considered an example of *implicit* or System 1 processing. Spellman, Holyoak, and Morrison (2001) demonstrated rapid priming based on semantic relations, using both naming and lexical decision paradigms. For example, participants were able to identify a related pair such as BEAR–CAVE as words more quickly when preceded (400 msec earlier) by BIRD–NEST (same relation) as opposed to BIRD–DESERT (unrelated). Note that although the pairs BIRD–NEST and BEAR–CAVE share the same relation (“lives in”), the objects themselves are not especially similar. Spellman et al. found that it was necessary to explicitly tell participants to pay attention to relations in order to obtain relation-based priming. However, such instructions were not essential in similar paradigms when the prime was presented for a longer duration and a relational judgment about the prime was required (Estes, 2003; Estes & Jones, 2006; see also Allen, Ibara, Seymour, Cordova, & Botvinick, 2010).

Relational priming may be especially potent when its application to objects of a certain type is overlearned. For example, Bassok, Chase, and Martin (1998) demonstrated that people are sensitive to what they termed *semantic alignment* of objects with the mathematical operation of addition. These investigators found that two sets of objects that each belong to the same general category (e.g., cats and dogs, both of which are animals) could readily be aligned with the addends of addition, whereas sets of objects that are functionally related (e.g., birds and cages) proved to be far less natural as addends. Extending this finding, Bassok, Pedigo, and Oskarsson (2008) showed that automatic activation of basic arithmetic facts (e.g., $3 + 4 = 7$) is modulated by prior presentation of aligned versus nonaligned word pairs (see also Fisher, Bassok, & Osterhout, 2010). The apparent automaticity of this phenomenon is consistent with people’s extensive experience in using the addition operation to solve problems involving a variety of semantically aligned object sets (e.g., blue and red marbles, cars and trucks, cupcakes and brownies). Importantly,

the priming effects observed in these studies did not in general aid in task performance, suggesting that priming was automatic.

Longer term priming may have contributed to the apparent unintended transfer effects observed in studies such as those of Schunn and Dunbar (1996) and Day and Goldstone (2011), as the source analog was itself processed deeply to solve a problem. Unintended transfer may not require developing a systematic mapping of the source to the target. Rather, more piecemeal transfer may occur based on activation of one or more key relational concepts (e.g., the concept of “inhibition” in the Schunn and Dunbar study).

Mapping

Mapping—the process of identifying corresponding elements of the source and target analogs—plays a central role in analogical inference. For meaningful situations such as problems or stories, adults with intact executive functions typically are able to establish correspondences based primarily on relational roles, even when direct similarity of mapped elements is low or uninformative (Gentner & Gentner, 1983; Gick & Holyoak, 1980, 1983).

ALIGNABILITY AND ATTENTIONAL FOCUS

Markman and Gentner (1993; for a review see Gentner & Markman, 1997) drew an important distinction between commonalities (the shared properties of mapped elements), alignable differences (differences between mapped elements), and nonalignable differences (differences between analogs involving unmapped elements). In Figure 13.4, for example, the car in the top picture can be mapped to the boat in the bottom picture based on their common roles (vehicles being towed). The car and boat also exhibit alignable differences. In contrast, the parking meter in the top picture has no clear corresponding element in the bottom picture, and hence it constitutes a nonalignable difference.

The basic impact of analogical mapping is to focus attention on the commonalities and (usually to a lesser extent) the alignable differences, while backgrounding the nonalignable differences. Gentner and Markman (1994) gave college students word pairs and asked them to list one difference each for as many pairs as possible under time pressure. The participants produced many more alignable than nonalignable differences. Contrary to

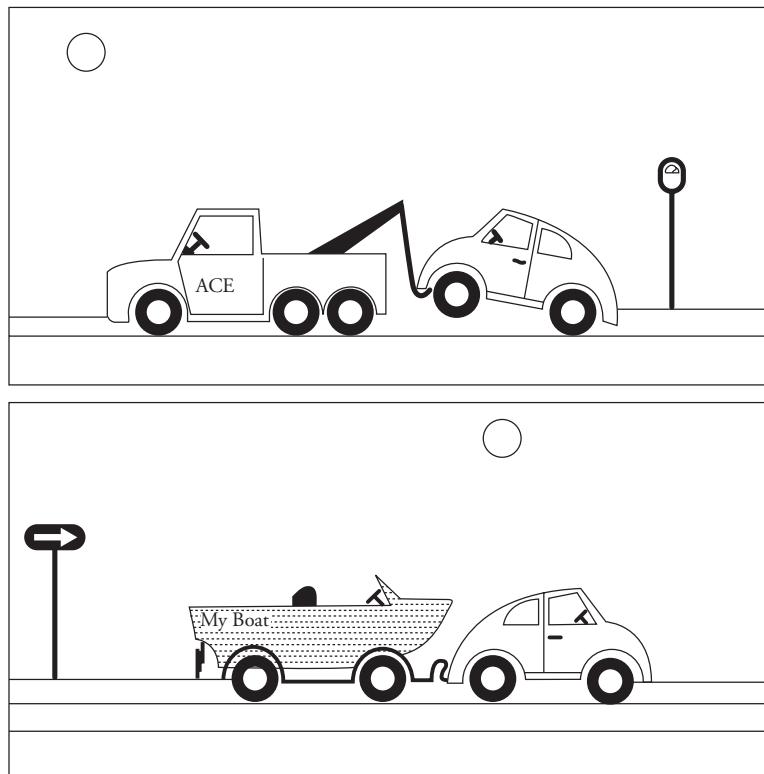


Fig. 13.4 Pictures illustrating types of analogy-based similarities and differences (Markman & Gentner, 1996). (Pictures courtesy of Art Markman.)

the commonsense idea that differences will be easier to find for dissimilar concepts, participants were actually more fluent in stating a difference for pairs of similar, alignable concepts (e.g., *hotel* – *motel*) than for dissimilar, nonalignable concepts (e.g., *kitten* – *magazine*), suggesting that the comparison process made the alignable differences especially salient. Relative to nonalignable differences, alignable differences have stronger effects on the perception of overall similarity (Markman & Gentner, 1996; see Goldstone & Son, Chapter 10), are more memorable (Markman & Gentner, 1997), and have a greater impact on choices made in a decision task (Markman & Medin, 1995).

COHERENCE IN ANALOGICAL MAPPING

The key idea of Holyoak and Thagard's (1989) multiconstraint theory of analogy is that several different kinds of constraints—similarity, structure, and purpose—all interact to determine the optimal set of correspondences between source and target. A good analogy is one that appears *coherent*, in the sense that multiple constraints converge on a solution that largely satisfies all constraints (Thagard, 2000). When constraints conflict, mappings may be ambiguous. For example, the two pictures shown in Figure 13.4 include a *cross-mapping*—the car in the top picture maps to the boat in the bottom picture on the basis of relational roles (both are vehicles being towed), but to the car in the bottom picture on the basis of direct similarity. Situations involving cross mappings are especially difficult, more so than analogies with less semantic overlap (Gentner & Toupin, 1986; Ross, 1989). Implicit cross mappings can also interfere with students' understanding of the intended interpretation of graphs and similar visuospatial representations (Gattis & Holyoak, 1996).

Comparisons based on perceptually rich stimuli, which afford an abundance of direct similarities between objects, typically lead to a lower frequency of relational responses relative to comparisons based on perceptually sparse stimuli (Markman & Gentner, 1993). In general, manipulations that increase attention to relations tend to encourage a relation-based response. For example, Markman and Gentner found that people who mapped three objects at once were more likely to map on the basis of similar relational roles than were people who mapped just one cross-mapped object, presumably because mapping multiple objects focuses greater attention on relations among them. Relational

language is an especially important influence on mapping. For preschool children, Loewenstein and Gentner (2005) found that explicitly describing a scene in terms of spatial relations increased the frequency of relation-based mappings (for a general discussion of language and thought, see Gleitman & Papafragou, Chapter 28).

In the absence of cross mappings, adults are generally able to integrate multiple constraints to establish coherent mappings, even for situations that are complex and somewhat ambiguous. For example, at the beginning of the first Gulf War in 1991, Spellman and Holyoak (1992) asked a group of American undergraduates a few questions to find out how they interpreted the analogy between the then-current situation in the Persian Gulf and World War II. The undergraduates were asked to suppose that Saddam Hussein, the President of Iraq, was analogous to Hitler (a popular analogy at the time). Regardless of whether they thought the analogy was appropriate, they were then asked to write down the most natural match in the World War II situation for various people and nations involved in the Gulf War, including the United States and its current President, George H. W. Bush. For those students who gave evidence that they knew the basic facts about World War II, the majority produced mappings that fell into one of two patterns, each coherent on relational grounds. Those students who mapped the United States to itself also mapped Bush to Franklin D. Roosevelt. Other students, in contrast, mapped the United States to Great Britain and Bush to Winston Churchill, the British Prime Minister (perhaps because Bush, like Churchill, led his nation and Western allies in early opposition to aggression). The analogy between the Persian Gulf situation and World War II thus generated a "bistable" mapping: People tended to provide mappings based on either of two coherent but mutually incompatible sets of correspondences.

Mapping is guided not only by relational structure and element similarity but also by the goals of the analogist (Holyoak, 1985). Particularly when the mapping is inherently ambiguous, the constraint of pragmatic centrality—relevance to goals—is critical (Holyoak, 1985). Spellman and Holyoak (1996) investigated the impact of processing goals on the mappings generated for inherently ambiguous analogies and found that mappings were predominately determined by those relations most relevant to the reasoner's goal.

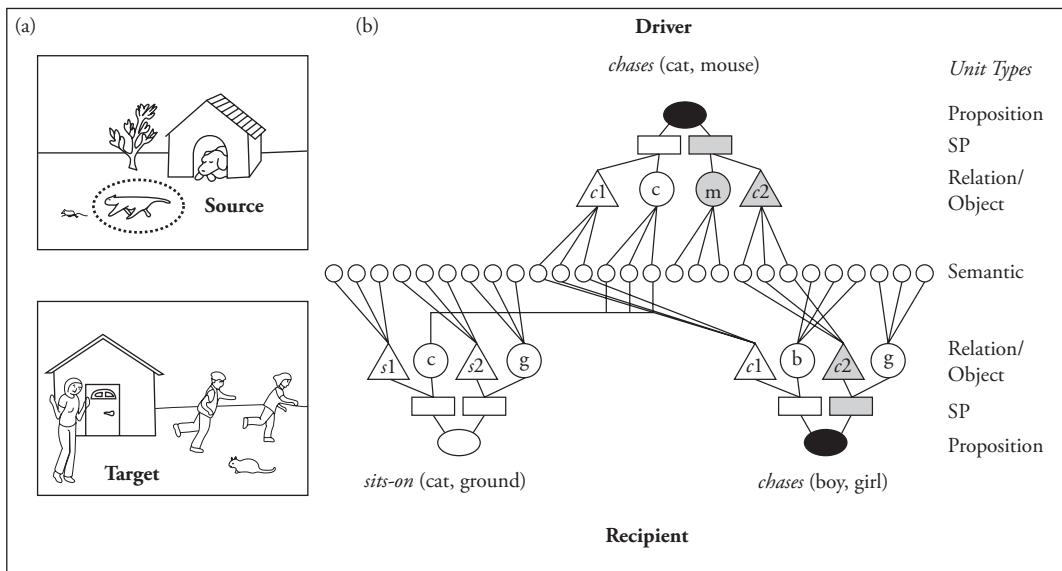


Fig. 13.5 (A) Example of one-relation/distractor scene-analogy problem (Richland et al., 2006); (B) LISA architecture as applied to this problem. In order for a reasoner to select the boy in the target as the correct analogical mapping to the cat in the source, units in the recipient representing the proposition *chases* (boy, girl) must inhibit corresponding units in the propositional structure containing the featureally similar “sitting cat” distractor. (Reprinted with permission from Morrison, Doumas, & Richland, 2011.)

DEVELOPMENTAL CHANGES IN ANALOGICAL MAPPING

Young children are particularly sensitive to direct similarity of objects. When asked to identify corresponding elements in two analogs when semantic and structural constraints conflict, their mappings are dominated by object similarity (Gentner & Toupin, 1986). The developmental transition toward greater reliance on relational structure in mapping has been termed the *relational shift* (Gentner & Rattermann, 1991). The empirical phenomenon of a relational shift is well established, but there has been some debate regarding the developmental mechanisms that may underlie it. Goswami and colleagues have argued that analogical reasoning is fundamentally available as a capacity from early infancy (and indeed, some analogical ability is apparent in 1-year-old children; Chen, Sanchez, & Campbell, 1997), but that children’s analogical performance increases with age due to the accretion of knowledge about relevant relations (Goswami, 1992, 2001; Goswami & Brown, 1989). Knowledge of relations is without doubt essential for analogical reasoning. Even for adults, expertise in a domain is a predictor of superior ability to process analogies in that domain (Novick & Holyoak, 1991).

However, although accretion of knowledge is certainly an important factor, there is now evidence

that maturational changes in cognitive functioning also drive developmental differences in analogical reasoning. These changes are associated with the maturation of the prefrontal cortex (see Diamond, 2002), which is not complete until adolescence. The prefrontal cortex underlies executive control (Diamond, 2006; see Morrison & Knowlton, Chapter 6), and in particular the capacity to manipulate complex information in working memory and to inhibit salient but task-inappropriate information and responses. A study by Richland, Morrison, and Holyoak (2006) illustrates how the need for working memory and inhibitory control influence the difficulty of analogical mapping at different ages. Figure 13.5a depicts an example of “scene-analogy” problems developed for use with children as young as 3–4 years old; Figure 13.5b illustrates how the LISA model (Hummel & Holyoak, 1997) would represent the mapping problem. For each pair of pictures, children were asked to identify the object in the bottom picture that “goes with” the object indicated by an arrow in the top picture. In some problems, such as that shown in Figure 13.5a, the child is confronted with a conflict between two possible answers, one relational and one based on perceptual and/or semantic similarity. The cat in the top picture perceptually resembles the cat in the bottom picture, but it plays a role (chasing a

mouse) that parallels the role played by the boy in the bottom picture (chasing a girl). Richland et al. found that young children were less likely to give the relational response when an alternative based on direct similarity was available. Inhibitory control is presumably required to avoid responding on the basis of direct similarity.

In addition, relational reasoning varies in its complexity, which has been linked to the number of relational roles relevant to an inference (Halford, 1993; Halford et al., 1998). The load on working memory will be less if a single relation is sufficient to determine the role-based inference (as in the example shown in Fig. 13.5a), compared to when multiple relations must be integrated to derive the inference. Richland et al. (2006) found that preschool children gave fewer relational responses when either a similar distractor was present in the bottom picture or when two relations had to be integrated. By age 13–14 years—roughly the age at which the prefrontal cortex has undergone substantial further maturation—children reliably gave the relational response even when multiple relations had to be integrated and a similar distractor was present. Children with autism, when matched to controls on measures of executive function, show comparable trends in analogical reasoning (Dawson et al., 2007; Morsanyi & Holyoak, 2010). This pattern of analogical development is consistent with what is known about the neural basis for analogical reasoning in adults, as discussed earlier.

Analogical Inference

Analogical inference—using a source analog to form a new conjecture, whether it be a step toward solving a math problem (see Bassok & Novick, Chapter 21), a scientific hypothesis (Dunbar & Klahr, Chapter 35), a basis for deciding a legal case (Spellman & Schauer, Chapter 36), or finding a diagnosis for puzzling medical symptoms (Patel et al., Chapter 37)—is the fundamental purpose of analogical reasoning (Bartha, 2010). Mapping serves to highlight correspondences between the source and target. These correspondences provide the input to an inference engine that generates new target propositions.

The basic algorithm for analogical inference used by all major computational models has been termed “copy with substitution and generation,” or CWSG (Holyoak, Novick, & Melz, 1994), and involves constructing target analogs based on unmapped source propositions by substituting the corresponding

target element (if known) for each source element, and if no corresponding target element is known, postulating one as needed. CWSG allows the generation of structured propositions about the target (as opposed to simple associations) because of its reliance on variable binding and mapping. In this key respect, inference by CWSG is similar to rule-based inferences of the sort modeled by production systems (e.g., Anderson & Lebiere, 1998; see Doumas & Hummel, Chapter 5; Koedinger & Roll, Chapter 40). However, the constraints on analogical mapping are more fluid than are the typical constraints on matching in a production system. CWSG is more flexible in that unlike production rules, there is no strict division between a “left-hand side” to be matched and a “right-hand side” that creates an inference. Rather, any subset of the two analogs may provide an initial mapping, and the unmapped remainder of the source may be used to create target inferences (giving rise to the property of *omnidirectional access* in analogical inference; Halford et al., 1998). Analogical inference might be described as a “strong weak method”—a domain-general method (see Bassok & Novick, Chapter 21) that can be extremely powerful if the requisite knowledge about a source analog is available (though useless if an appropriate source is lacking).

INTEGRATING ANALOGICAL INFERENCE WITH CAUSAL MODELS

The CWSG algorithm, and analogical inference in general, can fail in a variety of ways. If critical elements are difficult to map (e.g., because of strong representational asymmetries, such as those that hinder mapping a discrete set of elements to a continuous variable; Bassok & Olseth, 1995; Bassok & Holyoak, 1989), then no inferences can be constructed. If elements are mismapped, corresponding inference errors will result (Holyoak et al., 1994; Reed, 1987). Most important, the great fluidity of CWSG has its downside. Without additional constraints on when CWSG is invoked, *any* unmapped source proposition would generate an inference about the target. Such a loose criterion for inference generation would lead to rampant errors whenever the source was not isomorphic to a subset of the target; and such isomorphism will virtually never hold for problems of realistic complexity. Accordingly, additional constraints are required (Clement & Gentner, 1991; Holyoak et al., 1994; Markman, 1997).

Lassaline (1996) provided evidence of factors that appear to constrain analogical inferences. She had

college students read descriptions of the properties of hypothetical animals, and then rate various possible target inferences for the probability that the conclusion would be true, given the information in the premise. Participants rated potential inferences as more probable when the source and target analogs shared more attributes, and hence were more similar. In addition, the presence of a causal relation in the source made an inference more credible. For example, if the source and target animals were both described as having a weak immune system, and for the source the weak immune system was stated to “cause” an acute sense of smell, then the inference that the target animal also has an acute sense of smell would be bolstered relative to stating only that the source animal had a weak immune system “and” an acute sense of smell. The benefit conveyed by the linking relation was reduced if it was less clearly causal (“develops before”). Lassaline’s findings thus imply that although analogical inferences are influenced by the overall similarity of the analogs, causal relations in the source play an especially important role.

Work by Lee and Holyoak (2008) demonstrated the close connection between analogical inference and the operation of representations that have been termed *causal models* (Waldmann & Holyoak, 1992; see Cheng & Buehner, Chapter 12)—a network of cause-effect relations characterizing each analog. Holyoak, Lee, and Lu (2010) formalized this integration by extending a Bayesian model of causal learning (Lu et al., 2008) to deal with analogical inference. The basic idea is that for empirical analogies, the causal model of the source analog (including information about the strength distributions associated with individual causal links), coupled with the mapping of source to target, provide the input to CWSG. This procedure constrains CWSG is a way that favors accurate and useful inferences, generating as its output an elaborated causal model of the target. This causal model is then used to evaluate the probability of specific inferences about the target. By treating analogical and causal inference within a unifying theoretical framework, it proved possible to explain situations in which the strengths of inferences about the target are dissociable from the overall similarity of the source and target. Most dramatically, Lee and Holyoak showed that if the source exhibits an effect *despite* the presence of a preventive cause, then people judge the effect to be *more* likely in the target if it lacks the preventer (even though absence of the preventer reduces overall similarity of the source and target).

Rather than considering analogical inference in isolation, it is useful to view the entire transfer process as the joint product of causal learning and relational mapping: The reasoner learns the causal structure of the source, maps the source to target, applies CWSG to augment the causal model of the target, and then uses the resulting model to evaluate an open-ended range of potential inferences about the target. The model of the target generated on the basis of the source will often be imperfect, so that additional postanalogical processes of *adaptation* will be required to accommodate goal-relevant aspects of the target that are not predictable from the source (Carbonell, 1983; Holyoak et al., 1994).

ANALOGICAL INFERENCES AS “FALSE MEMORIES”

An important question concerns when analogical inferences are made, and how inferences relate to facts about the target analog that are stated directly. One extreme possibility is that people only make analogical inferences when instructed to do so, and that inferences are carefully “marked” as such, so that they will never be confused with known facts about the target. At the other extreme, it is possible that some analogical inferences are triggered when the target is first processed (given that the source has been activated), and that such inferences are then integrated with prior knowledge of the target. One paradigm for addressing this issue is based on testing for false “recognition” of potential inferences in a subsequent memory test. The logic of the recognition paradigm is that if an inference has been made and integrated with the rest of the target analog, then later the reasoner will believe that the inference had been directly presented, in effect having created a “false memory” (see Brainerd & Reyna, 2005).

Early work by Schustack and Anderson (1979) provided evidence that people sometimes falsely report that analogical inferences were actually presented as facts. Blanchette and Dunbar (2002) performed a series of experiments designed to assess when analogical inferences are made. They had college students (in Canada) read a text describing a current political issue, possible legalization of marijuana use, which served as the target analog. Immediately afterward, half the students read, “The situation with marijuana can be compared to...” followed by an additional text describing the period early in the 20th century when alcohol use was prohibited. Importantly, the students in the analogy condition were not told how prohibition mapped onto the

marijuana debate, nor were they asked to draw any inferences. After a delay (1 week in one experiment, 15 minutes in another), the students were given a list of sentences and were asked to decide whether each sentence had actually been presented in the text about marijuana use. The critical items were sentences such as “The government could set up agencies to control the quality and take over the distribution of marijuana.” These sentences had never been presented; however, they could be generated as analogical inferences by CWSG, based on a parallel statement contained in the source analog (“The government set up agencies to control the quality and take over the distribution of alcohol”). Blanchette and Dunbar found that students in the analogy condition said “yes” to analogical inferences about 50% of the time, whereas control subjects who had not read the source analog about prohibition said “yes” only about 25% of the time. This tendency to falsely “recognize” analogical inferences that had never been read was obtained both after long and short delays, and with both familiar and less familiar materials. Similar findings have been obtained by Perrott, Gentner, and Bodenhausen (2005).

It thus appears that when people notice the connection between a source and target, and they are sufficiently engaged in an effort to understand the target situation, analogical inferences will often be generated and then integrated with prior knowledge of the target. In some cases such transfer may be unintended and involve relational priming, as discussed earlier. At least sometimes, an analogical inference becomes accepted as a stated fact. Like relational priming, this is a case in which relational transfer does not necessarily improve performance of the target task (recognition memory). Such findings have important implications for understanding how analogical reasoning can operate as a tool for persuasion.

Relational Generalization and Schema Induction

In addition to generating local inferences about the target, analogical reasoning can give rise to relational generalizations—abstract schemas that establish an explicit representation of the commonalities between the source and target. *Comparison*—not simply passive accumulation of information about distributions of features across examples, but active generation of structural correspondences—lies at the heart of analogical reasoning. Comparison of multiple analogs can result not only in a specific mapping

but also in the induction of a schema, which in turn will facilitate subsequent transfer to additional analogs. The induction of such schemas has been demonstrated in both adults (Catrambone & Holyoak, 1989; Gick & Holyoak, 1983) and young children (Brown, Kane, & Echols, 1986; Chen & Daehler, 1989; Holyoak et al., 1984; Kotovsky & Gentner, 1996; Loewenstein & Gentner, 2001; Namy & Gentner, 2002).

Comparison has been shown to guide schema formation in teaching such complex topics as negotiation strategies (Loewenstein, Thompson, & Gentner, 1999, 2003; see Loewenstein, Chapter 38). There is also evidence that comparison may play a key role in learning role-based relations (e.g., comparative adjectives such as “bigger than”) from nonrelational inputs (Doumas et al., 2008), and in language learning more generally (Gentner, 2010; Gentner & Namy, 2006). An important refinement of the use of comparison as a training technique is to provide a series of comparisons ordered “easy to hard,” where the early pairs share salient surface similarities as well as less salient relational matches, and the later pairs share only relational matches. This “progressive alignment” strategy serves to promote a kind of analogical bootstrapping, using salient similarities to aid the learner in identifying appropriate mappings between objects that also correspond with respect to their relational roles (Kotovsky & Gentner, 1996).

FACTORS THAT INFLUENCE SCHEMA INDUCTION

People are able to induce schemas by comparing just two analogs to one another (Gick & Holyoak, 1983). Indeed, people will form schemas simply as a side effect of applying one solved source problem to an unsolved target problem (Novick & Holyoak, 1991; Ross & Kennedy, 1990). In the case of problem schemas, more effective schemas are formed when the goal-relevant relations are the focus rather than incidental details (Brown et al., 1986; Brown, Kane, & Long, 1989; Gick & Holyoak, 1983). In general, any kind of processing that helps people focus on the underlying functional structure of the analogs, thereby encouraging learning of more effective problem schemas, will improve subsequent transfer to new problems. For example, Gick and Holyoak (1983) found that induction of a “convergence” schema from two disparate analogs was facilitated when each story stated the underlying solution principle abstractly: “If you need a large

force to accomplish some purpose, but are prevented from applying such a force directly, many smaller forces applied simultaneously from different directions may work just as well." In some circumstances transfer can also be improved by having the reasoner generate a problem analogous to an initial example (Bernardo, 2001). Other work has shown that abstract diagrams that highlight the basic idea of using multiple converging forces can aid in schema induction and subsequent transfer (Beveridge & Parkins, 1987; Gick & Holyoak, 1983).

Although two examples can suffice to establish a useful schema, people are able to incrementally develop increasingly abstract schemas as additional examples are provided (Brown et al., 1986; Brown et al., 1989; Catrambone & Holyoak, 1989). Even with multiple examples that allow novices to start forming schemas, people may still fail to transfer the analogous solution to a problem drawn from a different domain if a substantial delay intervenes or if the context is changed (Spencer & Weisberg, 1986). Nonetheless, as novices continue to develop more powerful schemas, long-term transfer in an altered context can be dramatically improved (Barnett & Koslowski, 2002). For example, Catrambone and Holyoak (1989) gave college students a total of three convergence analogs to study, compare, and solve. The students were first asked a series of detailed questions designed to encourage them to focus on the abstract structure common to two of the analogs. After this abstraction training, the students were asked to solve another analog from a third domain (not the tumor problem), after which they were told the convergence solution to it (which most students were able to generate themselves). Finally, a week later, the students returned to participate in a different experiment. After the other experiment was completed, they were given the tumor problem to solve. Over 80% of participants came up with the converging-rays solution without any hint. As the novice becomes an expert, the emerging schema becomes increasingly accessible and is triggered by novel problems that share its structure (see Koedinger & Roll, Chapter 40). Deeper similarities have been constructed between analogous situations that fit the schema.

IMPACT OF SCHEMAS ON RELATIONAL PROCESSING

As schemas are acquired from examples, they in turn guide future analog retrieval, mapping, and inference. We have already seen how schema

induction can increase subsequent transfer to novel problems (e.g., Bassok & Holyoak, 1989; Gick & Holyoak, 1983; Loewenstein et al., 2003), as well as facilitate processing of metaphors (Gentner & Bowdle, 2008). In addition, a schema induced by comparing examples can work "backward" in memory, making it easier to retrieve analogous episodes (including autobiographical memories) that had been stored before the schema was acquired (Gentner, Loewenstein, Thompson, & Forbus, 2009).

Of course, schemas are often acquired through experience outside the laboratory. Such preexisting schemas can guide the interpretation of specific examples, thereby changing analogical mapping and inference. For example, Bassok, Wu, and Olseth (1995) examined analogical reasoning with algebra word problems similar to ones studied previously by Ross (1987, 1989). Participants were shown how to compute permutations using an example (the source problem) in which some items from a set of n members were randomly assigned to items from a set of m members (e.g., how many different ways can you assign three computers to three secretaries if there are n computers and m secretaries?). Participants were then tested for their ability to transfer this solution to new target problems. The critical basis for transfer hinged on the assignment relation in each analog—that is, what kinds of items (people or inanimate objects) served in the roles of n and m . In some problems (type OP) objects (n) were assigned to people (m ; e.g., "computers were assigned to secretaries"); in others (PO) people (n) were assigned to objects (m ; "secretaries were assigned to computers"). Solving a target problem required the participant to map the elements of the problem appropriately to the variables n and m in the equation for calculating the number of permutations.

Although all the problems were formally isomorphic, Bassok et al. (1995) demonstrated that people will typically interpret an "assign" relation between an object and a person as one in which the person *gets* the object. Importantly, the "get" schema favors interpreting the person as the recipient of the object no matter which entity occupies which role in the stated "assign" relation. These distinct interpretations of the stated "assign" statement yielded systematic consequences for analogical mapping. Given an OP source analog, Bassok et al. found that people tended to link the "assigned" object to the "received" role (rather than the "recipient" role) of the "get" schema, which in turn was then mapped

to the mathematical variable n , the number of “assigned” objects. As a result, when the target analog also had an OP structure, transfer was accurate (89%); but when the target was in the reversed PO structure, the object set continued to be linked to the “received” role of “get,” and hence erroneously mapped to n (0% correct!). Bassok et al.’s findings highlight the constructive and interactive nature of relational processing (see also Hofstadter & Mitchell, 1994).

Conclusions

Analogy is an important special case of role-based relational reasoning, a psychological process that generates inferences based on patterns of relational roles. At its core, analogy depends on comparison of situations. But humans do much more than just compare two analogs based on obvious similarities between their elements. Rather, analogical reasoning is a complex process of retrieving structured knowledge from long-term memory, representing and manipulating role-filler bindings in working memory, generating new inferences, and finding structured intersections between analogs to form new abstract schemas. For empirical analogies, analogical inference is guided by causal knowledge about how the source analog operates. Simpler types of relation-based transfer can be produced by relational priming.

Symbolic-connectionist models have the greatest promise in relating relational reasoning to its neural substrate. Human analogical reasoning is heavily dependent on working memory and other executive functions supported by the prefrontal cortex, with the frontopolar subregion being selectively activated when multiple relations must be integrated to solve a problem.

Future Directions

Computational Level

Theoretical work to date has specified qualitative constraints on analogical mapping and inference, often implemented in computer simulations. Recent efforts to integrate analogical inference with Bayesian causal models (Holyoak et al., 2010) suggest that human analogical inference may be approximately normative when the analogs can be represented as simple causal networks based on binary variables. However, a full computational-level analysis of relational reasoning using the modern Bayesian framework for induction (see Griffiths et al., Chapter 3) has not yet been offered. Given representations of

source and target analogs (including relevant prior knowledge), normative probability distributions for possible analogical mappings and inferences might in principle be derived. However, in practice this remains a challenging (perhaps even intractable) project, given that the types of relations involved in analogies are indefinitely diverse and include many different types of causal functions. One key requirement for applying the Bayesian framework to analogy will be greater theoretical integration of role-based relational representations with probabilistic inference.

Level of Representation and Algorithm

For over 30 years, a great deal of effort has been directed at the development of algorithmic models of analogical reasoning, formalized as computer simulations. In recent years, some of these models (based on the symbolic-connectionist framework) have begun to make contact with work on the neural substrate of relational reasoning. However, no model as yet comes close to providing a comprehensive account of how humans reason with relations.

One basic limitation is that human relational reasoning is far more flexible than any current simulation model (Bartha, 2010; Bassok et al., 1995; Hofstadter & Mitchell, 1994). The stage analysis typically used (for example, in this chapter) to organize analogical processing—retrieval, mapping, inference, schema induction (see Fig. 13.1)—is oversimplified. In everyday use of analogies, the entire process may be cyclic, interactive, and open ended. The effective representation of the source analog may be developed in the very process of reasoning with it. Multiple source analogs and schemas may be involved, some of them imagined rather than actual (as in the case of analogical “thought experiments”; see Bartha, 2010; Holyoak & Thagard, 1995). Typically there is no firm boundary around the information that counts as an individual analog. Causal knowledge, which is deeply embedded in the representation of individual cases, is almost inevitably linked to more general categories. Analogs, schemas, categories, and rules all interact with one another in the course of inductive inference (Holland et al., 1986). In addition, much more needs to be learned about how “full-blown” System 2 relational processing relates to simpler System 1 processing (e.g., relational priming).

A related limitation of current computational models of analogy is that their knowledge

representations typically must be hand-coded by the modeler, whereas human knowledge representations are formed autonomously. In effect, modelers have allowed themselves an indefinite number of free parameters to facilitate data-fitting. There have been recent efforts to extend analogy models to account for how humans learn basic perceptual relations from nonrelational inputs (Doumas et al., 2008). This is a welcome development, but even here, the nonrelational inputs have themselves been hand-coded by the modelers.

Closely related to the challenge of avoiding hand-coding of representations is the need to flexibly re-represent knowledge so as to render potential analogies perspicuous. Concepts often have a close conceptual relationship with more complex relational forms (Jackendoff, 1983). For example, causative verbs such as *lift* (e.g., “John lifted the hammer”) have very similar meanings to structures based on an explicit higher order relation, *cause* (e.g., “John caused the hammer to rise”). In such cases the causative verb serves as a “chunked” representation of a more elaborate predicate-argument structure. People are able to “see” analogies even when the analogs have very different linguistic forms (e.g., “John lifted the hammer in order to strike the nail” might be mapped onto “The Federal Reserve used an increase in interest rates as a tool in its efforts to drive down inflation”). A deeper understanding of human knowledge representation is a prerequisite for a complete theory of analogical reasoning.

Yet another limitation is that most research and modeling in the field of analogy has emphasized quasi-linguistic knowledge representations, but there is good reason to believe that reasoning in general has close connections to perception and action (see Hegarty & Stull, Chapter 31; Goldin-Meadow & Cook, Chapter 32). The ease of solving apparently isomorphic problems (e.g., isomorphs of the well-known Tower of Hanoi) can vary enormously depending on perceptual cues (Kotovsky & Simon, 1990). Models of analogy have not offered an adequate account of why the difficulty of solving problems and transferring solution methods to isomorphic problems is dependent on the difficulty of perceptually encoding key relations.

In addition, models of analogy have not been well integrated with models of problem solving (see Bassok & Novick, Chapter 21), despite the fact that analogy clearly affords an important mechanism for solving problems. In its general form, problem solving requires sequencing multiple operators, establishing

subgoals, and using combinations of rules to solve related but nonisomorphic problems. These basic requirements are beyond the capabilities of current computational models of analogy. The integration of analogy models with models of general problem solving remains an important research goal.

Neural Implementation

The recent advances in understanding the neural substrate of relational reasoning (in particular, the roles played by specific areas of prefrontal cortex operating within broader neural circuits) have set the stage for further work on how analogies are computed by the brain (see Morrison & Knowlton, Chapter 6). For example, tasks involving analogical processing, like those designed to elicit insight (see van Steenburgh et al., Chapter 24), should prove useful in investigating connections between the neural bases of cognition and emotion. Careful studies will be required to determine how the neural systems involved in analogical reasoning relate to those involved in other forms of role-based relational reasoning, such as deductive and linguistic inferences (e.g., Monti, Parsons, & Osherson, 2009).

Given that a significant evolutionary gap may separate human role-based relational reasoning from the capabilities of other extant primate species (Penn et al., 2008), animal models may not provide fully adequate models of human reasoning. Testing the mechanisms postulated by symbolic-connectionist models (Hummel & Holyoak, 1997), such as dynamic binding controlled by synchronous neural activity in the gamma band, and rapid cortical learning of mapping connections, will require noninvasive neuroimaging techniques that provide extremely fine temporal resolution coupled with good spatial resolution (but see Lu et al., 2006, for an example of how a behavioral priming technique may be useful for assessing the role of synchrony in perceptual representation). The emerging field of cognitive neurogenetics (see Green & Dunbar, Chapter 7) will doubtless provide deeper insights into the neural basis of human analogical reasoning.

Translational Research

In parallel with continued basic research on role-based relational reasoning, we can anticipate advances in many application areas, such as the use of analogies in teaching and learning (e.g., Richland, Stigler, & Holyoak, 2012), as aids to creative design, and in applications to computer-based search algorithms. The limits of analogical

applications are roughly coextensive with those of human imagination.

Acknowledgments

Preparation of this chapter was supported by grant N000140810186 from the Office of Naval Research, and grant R305C080015 from the Institute of Education Sciences. I thank Miriam Bassok, Alex Doumas, John Hummel, Art Markman, Bob Morrison, and Derek Penn for helpful discussions and suggestions.

References

- Allen, K., Ibara, S., Seymour, A., Cordova, N., & Botvinick, M. (2010). Abstract structural representations of goal-directed behavior. *Psychological Science*, 21, 1518–1524.
- Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.
- Asaad, W. F., Rainer, G., & Miller, E. K. (1998). Neural activity in the primate prefrontal cortex during associative learning. *Neuron*, 21, 1399–1407.
- Barnett, S. M., & Koslowski, B. (2002). Solving novel, ill-defined problems: Effects of type of experience and the level of theoretical understanding it generates. *Thinking and Reasoning*, 8, 237–267.
- Bartha, P. (2010). *By parallel reasoning: The construction and evaluation of analogical arguments*. New York: Oxford University Press.
- Bassok, M., Chase, V. M., & Martin, S. A. (1998). Adding apples and oranges: Alignment of semantic and formal knowledge. *Cognitive Psychology*, 35, 99–134.
- Bassok, M., & Holyoak, K. J. (1989). Interdomain transfer between isomorphic topics in algebra and physics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 153–166.
- Bassok, M., & Olseth, K. L. (1995). Object-based representations: Transfer between cases of continuous and discrete models of change. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 1522–1538.
- Bassok, M., Pedigo, S. F., & Oskarsson, A. T. (2008). Priming addition facts with semantic relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34, 343–352.
- Bassok, M., Wu, L. L., & Olseth, K. L. (1995). Judging a book by its cover: Interpretative effects of content on problem-solving transfer. *Memory and Cognition*, 23, 354–367.
- Bernardo, A. B. I. (2001). Principle explanation and strategic schema abstraction in problem solving. *Memory and Cognition*, 29, 627–633.
- Beveridge, M., & Parkins, E. (1987). Visual representation in analogical problem solving. *Memory and Cognition*, 15, 230–237.
- Black, M. (1962). *Models and metaphors*. Ithaca, NY: Cornell University Press.
- Blanchette, I., & Dunbar, K. (2000). Analogy use in naturalistic settings: The influence of audience, emotion, and goal. *Memory and Cognition*, 28, 730–735.
- Blanchette, I., & Dunbar, K. (2001). How analogies are generated: The roles of structural and superficial similarity. *Memory and Cognition*, 28, 108–124.
- Blanchette, I., & Dunbar, K. (2002). Representational change and analogy: How analogical inferences alter target representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 672–685.
- Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition*, 75, 1–28.
- Brainerd, C. J., & Reyna, V. F. (2005). *The science of false memory*. New York: Oxford University Press.
- Brown, A. L., Kane, M. J., & Echols, C. H. (1986). Young children's mental models determine analogical transfer across problems with a common goal structure. *Cognitive Development*, 1, 103–121.
- Brown, A. L., Kane, M. J., & Long, C. (1989). Analogical transfer in young children: Analogies as tools for communication and exposition. *Applied Cognitive Psychology*, 3, 275–293.
- Bunge, S. A., Wendelken, C., Badre, D., & Wagner, A. D. (2005). Analogical reasoning and prefrontal cortex: Evidence for separable retrieval and integration mechanisms. *Cerebral Cortex*, 15, 239–249.
- Carbonell, J. G. (1983). Learning by analogy: Formulating and generalizing plans from past experience. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (pp. 137–161). Palo Alto, CA: Tioga.
- Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, 97, 404–431.
- Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 1147–1156.
- Cattell, R. B. (1971). *Abilities: Their structure, growth, and action*. Boston, MA: Houghton-Mifflin.
- Chen, Z. (1996). Children's analogical problem solving: The effects of superficial, structural, and procedural similarity. *Journal of Experimental Child Psychology*, 62, 410–431.
- Chen, Z., & Daehler, M. W. (1989). Positive and negative transfer in analogical problem solving by 6-year-old children. *Cognitive Development*, 4, 327–344.
- Chen, Z., Sanchez, R. P., & Campbell, T. (1997). From beyond to within their grasp: The rudiments of analogical problem solving in 10- and 13-month-olds. *Developmental Psychology*, 33, 790–801.
- Cho, S., Holyoak, K. J., & Cannon, T. (2007). Analogical reasoning in working memory: Resources shared among relational integration, interference resolution, and maintenance. *Memory and Cognition*, 35, 1445–1455.
- Cho, S., Moody, T. D., Fernandino, L., Mumford, J. A., Poldrack, R. A., Cannon, T. D., ... Holyoak, K. J. (2010). Common and dissociable prefrontal loci associated with component mechanisms of analogical reasoning. *Cerebral Cortex*, 20, 524–533.
- Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J. K., Holyoak, K. J., & Gabrieli, J. D. E. (2001). Rostral-lateral prefrontal cortex involvement in relational integration during reasoning. *NeuroImage*, 14, 1136–1149.
- Clement, C. A., & Gentner, D. (1991). Systematicity as a selection constraint in analogical mapping. *Cognitive Science*, 15, 89–132.
- Cromer, J. A., Machon, M., & Miller, E. K. (2011). Rapid association learning in the primate prefrontal cortex in the absence of behavioral reversals. *Journal of Cognitive Neuroscience*, 23, 1823–1828.
- Dawson, M., Soulières, I., Gernsbacher, M. A., & Mottron, L. (2007). The level and nature of autistic intelligence. *Psychological Science*, 18, 657–662.

- Day, S. B., & Gentner, D. (2007). Nonintentional analogical inference in text comprehension. *Memory and Cognition*, 35, 39–49.
- Day, S. B., & Goldstone, R. L. (2011). Analogical transfer from a simulated physical system. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 551–567.
- Diamond, A. (2002). Normal development of prefrontal cortex from birth to young adulthood: Cognitive functions, anatomy, and biochemistry. In D. T. Stuss & R. T. Knight (Eds.), *Principles of frontal lobe function* (pp. 466–503). New York: Oxford University Press.
- Diamond, A. (2006). The early development of executive function. In E. Bialystok & F. I. M. Craik (Eds.), *Lifespan cognition: Mechanisms of change* (pp. 70–95). New York: Oxford University Press.
- Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, 115, 1–43.
- Duncan, J., Seitz, R. J., Kolodny, J., Bor, D., Herzog, H., Ahmed, A.,...Emslie, H. (2000). A neural basis for general intelligence. *Science*, 289, 457–460.
- Duncker, K. (1945). On problem solving. *Psychological Monographs*, 58 (Whole No. 270).
- Estes, Z. (2003). Attributive and relational processes in nominal combination. *Journal of Memory and Language*, 48, 304–319.
- Estes, Z., & Jones, L. L. (2006). Priming via relational similarity: A copper horse is faster when seen through a glass eye. *Journal of Memory and Language*, 55, 89–101.
- Evans, T. G. (1968). A program for the solution of geometric-analogy intelligence test questions. In M. Minsky (Ed.), *Semantic information processing* (pp. 271–353). Cambridge, MA: MIT Press.
- Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41, 1–63.
- Fauconnier, G. (2001). Conceptual blending and analogy. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 255–285). Cambridge, MA: MIT Press.
- Fauconnier, G., & Turner, M. (1998). Conceptual integration networks. *Cognitive Science*, 22, 133–187.
- Feldman, V., & Kokinov, B. (2009). Anxiety restricts the analogical search in an analogy generation task. In B. Kokinov, K. J. Holyoak, & D. Gentner (Eds.), *New frontiers in analogy research: Proceedings of the Second International Conference on Analogy* (pp. 117–126). Sofia, Bulgaria: New Bulgarian University Press.
- Fisher, K. J., Bassok, M., & Osterhout, L. (2010). When two plus two does not equal four: Event-related potential responses to incongruous arithmetic word problems. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1571–1576). Austin, TX: Cognitive Science Society.
- Gattis, M., & Holyoak, K. J. (1996). Mapping conceptual to spatial relations in visual reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 231–239.
- Gentner, D. (1982). Are scientific analogies metaphors? In D. S. Miall (Eds.), *Metaphor: Problems and perspectives* (pp. 106–132). Brighton, England: Harvester Press.
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155–170.
- Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, 34, 752–775.
- Gentner, D., & Bowdle, B. (2008). Metaphor as structure-mapping. In R. W. Gibbs, Jr. (Eds.), *Cambridge handbook of metaphor and thought* (pp. 109–128). New York: Cambridge University Press.
- Gentner, D., Falkenhainer, B., & Skorstad, J. (1988). Viewing metaphor as analogy. In D. H. Helman (Eds.), *Analogical reasoning: Perspectives of artificial intelligence, cognitive science, and philosophy* (pp. 171–177). Dordrecht, Netherlands: Kluwer.
- Gentner, D., & Gentner, D. R. (1983). Flowing waters or teeming crowds: Mental models of electricity. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 99–129). Hillsdale, NJ: Erlbaum.
- Gentner, D., Loewenstein, J., Thompson, L., & Forbus, K. D. (2009). Reviving inert knowledge: Analogical abstraction supports relational retrieval of past events. *Cognitive Science*, 33, 1343–1382.
- Gentner, D., & Markman, A. B. (1994). Structural alignment in comparison: No difference without similarity. *Psychological Science*, 5, 152–158.
- Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, 52, 45–56.
- Gentner, D., & Namy, L. L. (2006). Analogical processes in language learning. *Current Directions in Psychological Science*, 15, 297–301.
- Gentner, D., & Rattermann, M. (1991). Language and the career of similarity. In S. A. Gelman & J. P. Byrnes (Eds.), *Perspectives on thought and language: Interrelations in development* (pp. 225–277). London: Cambridge University Press.
- Gentner, D., Rattermann, M., & Forbus, K. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology*, 25, 524–575.
- Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science*, 10, 277–300.
- Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, 12, 306–355.
- Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, 15, 1–38.
- Glucksberg, S., Gildea, P., & Bookin, H. (1982). On understanding nonliteral speech: Can people ignore metaphors? *Journal of Verbal Learning and Verbal Behaviour*, 21, 85–98.
- Glucksberg, S., & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review*, 97, 3–18.
- Glucksberg, S., McClone, M. S., & Manfredi, D. (1997). Property attribution in metaphor comprehension. *Journal of Memory and Language*, 36, 50–67.
- Goode, M. R., Dahl, D. W., & Moreau, C. P. (2010). The effect of experiential analogies on consumer perception and attitudes. *Journal of Marketing Research*, 42, 274–286.
- Goswami, U. (1992). *Analogical reasoning in children*. Hillsdale, NJ: Erlbaum.
- Goswami, U. (2001). Analogical reasoning in children. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 437–470). Cambridge, MA: MIT Press.
- Goswami, U., & Brown, A. L. (1989). Melting chocolate and melting snowmen: Analogical reasoning and causal relations. *Cognition*, 35, 69–95.
- Green, A. E., Fugelsang, J. A., Kraemer, D. J., Shamosh, N. A., & Dunbar, K. N. (2006). Frontopolar cortex mediates abstract integration in analogy. *Brain Research*, 1096, 125–137.

- Green, A. E., Kraemer, D. J., Fugelsang, A. J., Gray, J. R., & Dunbar, K. N. (2010). Connecting long distance: Semantic distance in analogical reasoning modulates frontopolar cortex activity. *Cerebral Cortex*, 20, 70–76.
- Halford, G. S. (1993). *Children's understanding: The development of mental models*. Hillsdale, NJ: Erlbaum.
- Halford, G. S., & Wilson, W. H. (1980). A category theory approach to cognitive development. *Cognitive Psychology*, 12, 356–411.
- Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, 21, 803–831.
- Halford, G. S., Wilson, W. H., & Phillips, S. (2010). Relational knowledge: The foundation of higher cognition. *Trends in Cognitive Sciences*, 14, 497–505.
- Hesse, M. (1966). *Models and analogies in science*. Notre Dame, IN: Notre Dame University Press.
- Hofstadter, D. R., & Mitchell, M. (1994). The copycat project: A model of mental fluidity and analogy-making. In K. J. Holyoak & J. A. Barnden (Eds.), *Analogical connections. Advances in connectionist and neural computation theory* (Vol. 2, pp. 31–112). Norwood, NJ: Ablex.
- Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press.
- Holyoak, K. J. (1982). An analogical framework for literary interpretation. *Poetics*, 11, 105–126.
- Holyoak, K. J. (1985). The pragmatics of analogical transfer. In G. H. Bower (Ed.), *The psychology of learning and motivation*, (Vol. 19, pp. 59–87). New York: Academic Press.
- Holyoak, K. J., Junn, E. N., & Billman, D. O. (1984). Development of analogical problem-solving skill. *Child Development*, 55, 2042–2055.
- Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory and Cognition*, 15, 332–340.
- Holyoak, K. J., Lee, H. S., & Lu, H. (2010). Analogical and category-based inference: A theoretical integration with Bayesian causal models. *Journal of Experimental Psychology: General*, 139, 702–727.
- Holyoak, K. J., Novick, L. R., & Melz, E. R. (1994). Component processes in analogical transfer: Mapping, pattern completion, and adaptation. In K. J. Holyoak & J. A. Barnden (Eds.), *Advances in connectionist and neural computation theory. Vol. 2: Analogical connections* (pp. 113–180). Norwood, NJ: Ablex.
- Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science*, 13, 295–355.
- Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought*. Cambridge, MA: MIT Press.
- Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104, 427–466.
- Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, 110, 220–263.
- Hunt, E. B. (1974). Quote the raven? Nevermore! In L. W. Gregg (Ed.), *Knowledge and cognition* (pp. 129–157). Hillsdale, NJ: Erlbaum.
- Jackendoff, R. (1983). *Semantics and cognition*. Cambridge, MA: MIT Press.
- Keane, M. T. (1987). On retrieving analogues when solving problems. *Quarterly Journal of Experimental Psychology*, 39A, 29–41.
- Keane, M. T. (1997). What makes an analogy difficult? The effects of order and causal structure on analogical mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 946–967.
- Keane, M. T., & Brayshaw, M. (1988). The incremental analogical machine: A computational model of analogy. In D. Sleeman (Ed.), *European working session on learning* (pp. 53–62). London: Pitman.
- Keysar, B. (1989). On the functional equivalence of literal and metaphorical interpretations in discourse. *Journal of Memory and Language*, 28, 375–385.
- Khong, Y. F. (1992). *Analogies at war: Korea, Munich, Dien Bien Phu, and the Vietnam decisions of 1965*. Princeton, NJ: Princeton University Press.
- Knowlton, B. J., & Holyoak, K. J. (2009). Prefrontal substrate of human relational reasoning. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (4th ed., pp. 1005–1017). Cambridge, MA: MIT Press.
- Kokinov, B. N., & Petrov, A. A. (2001). Integration of memory and reasoning in analogy-making: The AMBR model. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 59–124). Cambridge, MA: MIT Press.
- Kolodner, J. L. (1983). Reconstructive memory: A computer model. *Cognitive Science*, 7, 281–328.
- Kolodner, J. L. (1993). *Case-based reasoning*. San Mateo, CA: Morgan Kaufmann.
- Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, 67, 2797–2822.
- Kotovsky, K., & Simon, H. A. (1990). What makes some problems really hard? Explorations in the problem space of difficulty. *Cognitive Psychology*, 22, 143–183.
- Krawczyk, D. C., Morrison, R. G., Viskontas, I., Holyoak, K. J., Chow, T. W., Mendez, M. F., ... Knowlton, B. J. (2008). Distraction during relational reasoning: The role of prefrontal cortex in interference control. *Neuropsychologia*, 46, 2020–2032.
- Kroger, J. K., Saab, F. W., Fales, C. L., Bookheimer, S. Y., Cohen, M. S., & Holyoak, K. J. (2002). Recruitment of anterior dorsolateral prefrontal cortex in human reasoning: A parametric study of relational complexity. *Cerebral Cortex*, 12, 477–485.
- Kubose, T. T., Holyoak, K. J., & Hummel, J. E. (2002). The role of textual coherence in incremental analogical mapping. *Journal of Memory and Language*, 47, 407–435.
- Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago, IL: University of Chicago Press.
- Lakoff, G., & Turner, M. (1989). *More than cool reason: A field guide to poetic metaphor*. Chicago, IL: University of Chicago Press.
- Lassaline, M. E. (1996). Structural alignment in induction and similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 754–770.
- Lee, H. S., & Holyoak, K. J. (2008). The role of causal models in analogical inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34, 1111–1122.
- Loewenstein, J., & Gentner, D. (2001). Spatial mapping in preschoolers: Close comparisons facilitate far mappings. *Journal of Cognition and Development*, 2, 189–219.
- Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive Psychology*, 50, 315–353.

- Loewenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin and Review*, 6, 586–597.
- Loewenstein, J., Thompson, L., & Gentner, D. (2003). Analogical learning in negotiation teams: Comparing cases promotes learning and transfer. *Academy of Management Learning and Education*, 2, 119–127.
- Lu, H., Morrison, R. G., Hummel, J. E., & Holyoak, K. J. (2006). Role of gamma-band synchronization in priming of form discrimination for multiobject displays. *Journal of Experimental Psychology: Human Perception and Performance*, 32, 610–617.
- Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, 115, 955–982.
- Markman, A. B. (1997). Constraints on analogical inference. *Cognitive Science*, 21, 373–418.
- Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, 23, 431–467.
- Markman, A. B., & Gentner, D. (1996). Commonalities and differences in similarity comparisons. *Memory and Cognition*, 24, 235–249.
- Markman, A. B., & Gentner, D. (1997). The effects of alignability on memory. *Psychological Science*, 8, 363–367.
- Markman, A. B., & Loewenstein, J. (2010). Structural comparison and consumer choice. *Journal of Consumer Psychology*, 20, 126–137.
- Markman, A. B., & Medin, D. L. (1995). Similarity and alignment in choice. *Organizational Behavior and Human Decision Processes*, 63, 117–130.
- Markman, A. B., Taylor, E., & Gentner, D. (2007). Auditory presentation leads to better analogical retrieval than written presentation. *Psychonomic Bulletin and Review*, 14, 1101–1106.
- Markman, A. B., & Stilwell, C. H. (2001). Role-governed categories. *Journal of Experimental and Theoretical Artificial Intelligence*, 13, 329–358.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: W. H. Freeman.
- Medin, D. L., & Ross, B. H. (1989). The specific character of abstract thought: Categorization, problem-solving, and induction. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 5, pp. 179–195). New York: Cambridge University Press.
- Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24, 167–202.
- Minsky, M. (1975). A framework for representing knowledge. In P. Winston (Ed.), *The psychology of computer vision* (pp. 211–281). New York: McGraw-Hill.
- Monti, M. M., Parsons, L. M., & Osherson, D. N. (2009). The boundaries of language and thought in deductive inference. *Proceedings of the National Academy of Sciences USA*, 106, 12554–12559.
- Morrison, R. G., Doumas, L. A. A., & Richland, L. E. (2011). A computational account of children's analogical reasoning: Balancing inhibitory control in working memory and relational representation. *Developmental Science*, 14, 516–529.
- Morrison, R. G., Krawczyk, D. C., Holyoak, K. J., Hummel, J. E., Chow, T. W., Miller, B. L., & Knowlton, B. J. (2004). A neurocomputational model of analogical reasoning and its breakdown in frontotemporal lobar degeneration. *Journal of Cognitive Neuroscience*, 16, 260–271.
- Morsanyi, K., & Holyoak, K. J. (2010). Analogical reasoning ability in autistic and typically-developing children. *Developmental Science*, 13, 578–587.
- Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology*, 12, 252–284.
- Namy, L. L., & Gentner, D. (2002). Making a silk purse out of two sow's ears: Young children's use of comparison in category learning. *Journal of Experimental Psychology: General*, 131, 5–15.
- Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14, 510–520.
- Novick, L. R., & Holyoak, K. J. (1991). Mathematical problem solving by analogy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17, 398–415.
- Pask, C. (2003). Mathematics and the science of analogies. *American Journal of Physics*, 71, 526–534.
- Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and non-human minds. *Behavioral and Brain Sciences*, 31, 109–178.
- Perrott, D. A., Gentner, D., & Bodenhausen, G. V. (2005). Resistance is futile: The unwitting insertion of analogical inferences in memory. *Psychonomic Bulletin and Review*, 12, 696–702.
- Raven, J. C. (1938). *Progressive matrices: A perceptual test of intelligence, individual form*. London: Lewis.
- Reed, S. K. (1987). A structure-mapping model for word problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 124–139.
- Reitman, W. (1965). *Cognition and thought*. New York: Wiley.
- Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology*, 94, 249–271.
- Richland, L. E., Stigler, J. W., & Holyoak, K. J. (2012). Teaching the conceptual structure of mathematics. *Educational Psychologist*.
- Richland, L. E., Zur, O., & Holyoak, K. J. (2007). Cognitive supports for analogy in the mathematics classroom. *Science*, 316, 1128–1129.
- Robin, N., & Holyoak, K. J. (1995). Relational complexity and the functions of prefrontal cortex. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 987–997). Cambridge, MA: MIT Press.
- Ross, B. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 629–639.
- Ross, B. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 456–468.
- Ross, B. H., & Kennedy, P. T. (1990). Generalizing from the use of earlier examples in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 42–55.
- Schank, R. C. (1982). *Dynamic memory*. New York: Cambridge University Press.
- Schunn, C. D., & Dunbar, K. (1996). Priming, analogy, and awareness in complex reasoning. *Memory and Cognition*, 24, 271–284.
- Schustack, M. W., & Anderson, J. R. (1979). Effects of analogy to prior knowledge on memory for new information. *Journal of Verbal Learning and Verbal Behavior*, 18, 565–583.

- Seifert, C. M., McKoon, G., Abelson, R. P., & Ratcliff, R. (1986). Memory connections between thematically similar episodes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12, 220–231.
- Smith, E. E., Langston, C., & Nisbett, R. E. (1992). The case for rules in reasoning. *Cognitive Science*, 16, 1–40.
- Snow, R. E., Kyllonen, C. P., & Marshalek, B. (1984). The topography of ability and learning correlations. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (pp. 47–103). Hillsdale, NJ: Erlbaum.
- Spearman, C. (1923). *The nature of intelligence and the principles of cognition*. London: MacMillan.
- Spearman, C. (1927). *The abilities of man*. New York: Macmillan.
- Spearman, C. (1946). Theory of a general factor. *British Journal of Psychology*, 36, 117–131.
- Spellman, B. A., & Holyoak, K. J. (1992). If Saddam is Hitler then who is George Bush? Analogical mapping between systems of social roles. *Journal of Personality and Social Psychology*, 62, 913–933.
- Spellman, B. A., & Holyoak, K. J. (1996). Pragmatics in analogical mapping. *Cognitive Psychology*, 31, 307–346.
- Spellman, B. A., Holyoak, K. J., & Morrison, R. G. (2001). Analogical priming via semantic relations. *Memory and Cognition*, 29, 383–393.
- Spencer, R. M., & Weisberg, R. W. (1986). Context-dependent effects on analogical transfer. *Memory and Cognition*, 14, 442–449.
- Sternberg, R. J. (1977). Component processes in analogical reasoning. *Psychological Review*, 84, 353–378.
- Taylor, E. G., & Hummel, J. E. (2009). Finding similarity in a model of relational reasoning. *Cognitive Systems Research*, 10, 229–239.
- Thagard, P. (2000). *Coherence in thought and action*. Cambridge, MA: MIT Press.
- Thagard, P., & Shelley, C. (2001). Emotional analogies and analogical inference. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 335–362). Cambridge, MA: MIT Press.
- Tohill, J. M., & Holyoak, K. J. (2000). The impact of anxiety on analogical reasoning. *Thinking and Reasoning*, 6, 27–40.
- Tunteler, E., & Resing, W. C. M. (2002). Spontaneous analogical transfer in 4-year-olds: A microgenetic study. *Journal of Experimental Child Psychology*, 83, 149–166.
- Viskontas, I. V., Morrison, R. G., Holyoak, K. J., Hummel, J. E., & Knowlton, B. J. (2004). Relational integration, inhibition and analogical reasoning in older adults. *Psychology and Aging*, 19, 581–591.
- Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, 121, 222–236.
- Waltz, J. A., Knowlton, B. J., Holyoak, K. J., Boone, K. B., Mishkin, F. S., de Menezes Santos, M., . . . Miller, B. L. (1999). A system for relational reasoning in human prefrontal cortex. *Psychological Science*, 10, 119–125.
- Waltz, J. A., Lau, A., Grewal, S. K., & Holyoak, K. J. (2000). The role of working memory in analogical mapping. *Memory and Cognition*, 28, 1205–1212.
- Wendelken, C., Nakahbenko, D., Donohue, S. E., Carter, C. S., & S. A. Bunge, S. A. (2008). “Brain is to thought as stomach is to?” Investigating the role of rostral-lateral prefrontal cortex in relational reasoning. *Journal of Cognitive Neuroscience*, 20, 682–693.
- Wharton, C. M., Holyoak, K. J., Downing, P. E., Lange, T. E., Wickens, T. D., & Melz, E. R. (1994). Below the surface: Analogical similarity and retrieval competition in reminding. *Cognitive Psychology*, 26, 64–101.
- Wharton, C. M., Holyoak, K. J., & Lange, T. E. (1996). Remote analogical reminding. *Memory and Cognition*, 24, 629–643.
- Wharton, C. M., & Lange, T. E. (1994). Analogical transfer through comprehension and priming. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 934–939). Hillsdale, NJ: Erlbaum.
- Winston, P. H. (1980). Learning and reasoning by analogy. *Communications of the ACM*, 23, 689–703.
- Wolff, P., & Gentner, D. (2000). Evidence for role-neutral initial processing of metaphors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 529–541.

Explanation and Abductive Inference

Tania Lombrozo

Abstract

Everyday cognition reveals a sophisticated capacity to seek, generate, and evaluate explanations for the social and physical worlds around us. Why are we so driven to explain, and what accounts for our systematic explanatory preferences? This chapter reviews evidence from cognitive psychology and cognitive development concerning the structure and function of explanations, with a focus on the role of explanations in learning and inference. The findings highlight the value of understanding explanation and abductive inference both as phenomena in their own right and for the insights they provide concerning foundational aspects of human cognition, such as representation, learning, and inference.

Key Words: explanation, understanding, abduction, abductive inference, inference to the best explanation, self-explanation

Introduction

This chapter considers a ubiquitous but under-studied aspect of human cognition: explanation. Both children and adults seek explanations constantly. We wonder why events unfold in particular ways, why objects have specific properties, and why others behave as they do. Moreover, we have strong intuitions about what requires explanation and what counts as an adequate answer. Where do these intuitions come from? Why are we so driven to explain the social and physical worlds around us? And more generally, what is the role of explanation in cognition?

While there is no consensus on how to address these questions, research in cognitive development and cognitive psychology is beginning to make headway in finding answers (Keil, 2006; Keil & Wilson, 2000; Lombrozo, 2006; Wellman, 2011). These new developments complement large but focused literatures in social psychology (e.g., Anderson, Krull, & Weiner, 1996; Heider, 1958; Malle, 2004),

education (e.g., Chi, 2000; Roscoe & Chi, 2008), and philosophy (e.g., Salmon, 1989; Woodward, 2010), which have respectively emphasized explanations for human behavior, explanations in pedagogy, and explanations in science. Explanation has also been of interest in other cognitive science disciplines, such as artificial intelligence, where the generation of explanations has been proposed as a mechanism for learning (e.g., DeJong, 2004).

In this chapter, I focus on recent research in cognitive psychology and cognitive development. I begin by motivating the study of explanation, what explanations are, and their importance for understanding cognition. This is followed by more extensive discussions of the roles of explanation in learning and inference. While explanations and their effects are quite diverse, the findings support a common picture according to which the process of explaining recruits prior beliefs and a host of explanatory preferences, such as unification and simplicity, that jointly constrain subsequent processing. I conclude

by considering the current status of research on explanation and highlighting promising directions for future research.

Why Explanation Matters for Cognition

The study of explanation is motivated both by everyday experience and by theoretical considerations. Seeking, generating, and evaluating explanations is a central human preoccupation from an early age, and anecdotal evidence suggests that explanations play an important role in motivating discovery, communicating beliefs, and furthering understanding. Much of the recent interest in explanation, however, stems from theoretical and empirical developments that place explanation at the core of claims about conceptual representation, learning, and inference.

Since at least the 1980s, a prominent strand of thinking within cognitive science has postulated structured mental representations that share many of the properties of scientific theories (e.g., Carey, 1985; Gopnik & Meltzoff, 1997; Murphy & Medin, 1985; Wellman & Gelman, 1992; see Gelman & Frazier, Chapter 26). These so-called naïve, folk, or intuitive theories are semicoherent bodies of belief that support prediction, explanation, and intervention for a specified domain of application, such as physics or psychology. Almost all advocates for this approach have appealed to explanation in defining intuitive theories. For example, theories have been identified with “any of a host of mental explanations” (Murphy & Medin, 1985, p. 290) and characterized in terms of “laws and other explanatory mechanisms” (Carey, 1985, p. 201). It is clear, then, that an adequate account of intuitive theories rests on an adequate account of explanation (see also Lombrozo, 2010a; Machery, 2009; Margolis, 1995).

In parallel with a focus on intuitive theories, accounts of concepts and categorization have increasingly emphasized “explanation-based” reasoning in contrast to purely statistical or similarity-based reasoning (see Rips et al., Chapter 11). For example, categorization and category learning have been described as “special cases of inference to the best explanation” (Rips, 1989, p. 53), with a concept invoked “when it has a sufficient explanatory relation to an object” (Murphy & Medin, 1985, p. 295). And in tasks such as property generalization, deviations from patterns of reasoning based on similarity are often explained by appeal to causal or explanatory beliefs (e.g., Murphy & Allopenna, 1994; Rehder, 2006), pointing to the need for an

account of what makes some beliefs “explanatory,” and how those beliefs influence reasoning.

Finally, there is increasing evidence that prior beliefs significantly impact how reasoners learn and draw inferences. Generating and evaluating explanations may be mechanisms by which prior beliefs are brought to bear on a given task (Lombrozo, 2006), especially in knowledge-rich domains. The study of explanation is therefore not only of value in its own right, but as a window onto foundational aspects of cognition ranging from conceptual representation to learning and inference.

Defining Explanation

A natural place to begin the study of explanation is with a clear characterization of the object of study. In other words, what *is* explanation? Unfortunately, this innocuous question has proven quite complex. Some researchers define explanations as answers to why- or how-questions (e.g., Wellman, 2011), others as judgments about why an outcome occurred (e.g., Krull & Anderson, 2001), yet others as hypotheses positing causes that have what is being explained as their effects (e.g., Thagard, 2000; see also Einhorn & Hogarth, 1986). In all likelihood, explanation is not unitary: Research on explanation almost certainly spans multiple kinds of judgments and distinct cognitive mechanisms. “Explanation” is likely to be a family-resemblance term, picking out a cluster of related phenomena.

A first step toward precision, if not definition, is to distinguish explanation as a *product* from explanation as a *process* (see also Chin-Parker & Bradner, 2010). As a product, an explanation is a proposition or judgment, typically linguistic, that addresses an explicit or implicit request for an explanation. As a process, explanation is a cognitive activity that aims to generate one or more explanation “products” but need not succeed in order to be engaged. Most theories of explanation from philosophy are about explanations as products, whereas empirical research has been more mixed, with some research focusing on the characteristics of the product, and other research on the characteristics and consequences of the process.

A second useful distinction is between a “complete” explanation and a “selected” explanation (for related distinctions see Goodman et al., 2006; Railton, 1981). To illustrate, consider an explanation for Carl’s decision to order a slice of carrot cake over chocolate cake. In most contexts, noting a stable preference (“carrot cake is Carl’s favorite”) can

be sufficient to explain the decision, but this explanation rests on a variety of background assumptions (for example, that a preference for carrot cake would lead someone to choose carrot cake over chocolate cake), as well as a much deeper causal structure that could potentially be traced back to the origins of the universe. A selected explanation will correspond to the small subset of this explanatory structure that is identified as “the” explanation in a given context: probably the cake preference, but probably not the big bang.

This distinction is important in understanding explanation as both product and process. When considering explanation as a product, there is the selected explanation itself—what an individual might offer in response to a why-question—as well as a more complete explanation, which includes the cognitive infrastructure that supports the selected explanation. Of course, when it comes to human cognition, the complete explanation could be quite incomplete—it’s likely that human explanatory practice proceeds in the absence of fully specified explanations, even if there is always a more complete explanation that underlies the selected explanation (Keil, 2003; Rozenblit & Keil, 2002). When considering explanation as a process, it is useful to distinguish two distinct (although not necessarily independent) inferential steps: one to the complete explanation, and a second “selection process” to identify the selected explanation (which could, in principle, be equivalent to the complete explanation).

Potential insight into how selection occurs comes from the observation that explanations are typically contrastive: They account for one state of affairs in contrast to another (see e.g., Chin-Parker & Bradner, 2010; Garfinkel, 1990; Hart & Honore, 1985; Hilton, 1996; Hilton & Slugoski, 1986; Mackie, 1980; van Fraassen, 1980). The role of an implicit contrast is illustrated by an explanation attributed to Willie Sutton, a notorious bank robber. When asked why he robbed banks, he explained: “Because that’s where the money is.” The response is humorous in part because it addresses an inappropriate contrast: Sutton has explained why he robs banks *rather than robbing other places*, but he has failed to adequately answer the question most likely intended, namely why he robs banks *rather than not robbing banks*. The contrast provides a constraint on what should figure in a selected explanation: Explanations typically identify conditions that differentiate what is being explained (the explanans) from a counterfactual case in which the contrast

holds. Aspects common to the explanans and to the contrast likely figure in the complete explanation, which supports the selected explanation but need not be found explanatory. For example, Carl’s choice of carrot cake over chocolate cake is poorly explained (if at all) by noting that “cake is delicious,” as this piece of background does not explain why he chose carrot cake *as opposed to chocolate cake*. Recognizing the contrastive nature of explanation also reveals that a full specification of what is being explained involves identifying assumptions that are implicit in an explanation request.

The distinctions between explanation as product and process and complete versus selected explanations highlight one of the challenges in the study of explanation: providing cohesive theories that are nonetheless sensitive to this diversity. Explanatory products and processes are clearly related and mutually constraining, as are complete and selected explanations. Yet a failure to recognize these distinctions can make it difficult to relate different strands of research or to appreciate their implications for cognition.

The Structure of Explanations

Within philosophy, there have been several systematic attempts to provide definitions or precise specifications of what counts as an explanation (the product), and in particular the relationship that must obtain between an explanation and what it explains. Although typically intended as normative claims about explanation in science, these theories provide a useful starting point for psychological investigation and help identify the space of possible approaches. Initiating the contemporary study of explanation in philosophy of science, Hempel and Oppenheim (1948) proposed the deductive-nomological (DN) account of explanation, according to which an explanation involves initial conditions and general laws from which one can deduce what is being explained (the explanandum). This account was later extended to address inductive rather than deductive cases (the inductive-statistical [IS] account; see Hempel, 1965), but it still faced serious problems that led most philosophers to alternative accounts (see Salmon, 1989, for a review).

Currently, there are three popular approaches to explanation within philosophy of science. The two most prominent are causal theories (see Buehner & Cheng, Chapter 12), according to which explanations identify the cause(s) for an event or property (e.g., Salmon 1984; Woodward, 2003), and

subsumption or unification theories, which claim that an explanation subsumes the explanandum under an explanatory pattern, which can but need not be a law or counterfactual-supporting generalization (e.g., Friedman, 1974; Kitcher, 1989). These two approaches have also been fruitfully combined. For example, Strevens (2008) proposes a causal theory of explanation but relies on criteria involving unification to solve the selection problem of identifying a selected explanation from a complete one. A third approach, related to causal approaches, is to identify explanations with a description of causal mechanisms, where a mechanism consists of components, operations, and their (often hierarchical) organization, which realize what is being explained (e.g., Bechtel, 2008; Craver, 2007; Darden, 2006).

While many everyday and scientific explanations in fact satisfy all of these theories, the approaches differ in their account of *why* the explanation is explanatory. For example, consider explaining a disease by appeal to a virus. According to a causal theory, an occurrence of the disease can be explained by appeal to the presence of the virus because the virus causally accounts for the occurrence of the disease. According to a subsumption/unification theory, the virus explains the disease by subsuming the instance being explained under a broader generalization—that the particular virus leads to that particular disease—which in turn conforms to a broader generalization about viruses causing diseases, and so on to potentially greater levels of abstraction. While multiple theories can thus accommodate a wide range of common cases, they face distinct challenges. In general, causal and causal mechanism approaches have a difficult time handling cases of explanation that do not seem to involve causation at all, such as mathematical explanations. In contrast, subsumption/unification theories face the problem of specifying which kinds of patterns or unifications count as explanatory. A strategy potentially amenable to both causal and subsumption theories is to identify some kind of higher order, asymmetric dependence relationship or generative process that can subsume both causal and mathematical or formal relationships, and serve as a foundation for a theory of explanation with broader scope.

Causal, subsumption/unification, and mechanism-based approaches all find some empirical support within psychology. Many investigations of explanation have been restricted to causal cases, but some accounts of the effects of explanation on learning appeal to subsumption and unification (see

later section on “Explanation and Learning”). There is also evidence that knowledge of mechanisms is often invoked in explanations and constrains inference (e.g., Ahn & Bailenson, 1996; Ahn, Kalish, Medin, & Gelman, 1995; Koslowski, 1996; Shultz, 1980). However, no studies (to my knowledge) have attempted to differentiate these theories empirically by isolating cases for which the theories generate different predictions.

The Functions of Explanation

Intertwined with questions about the structure of explanations are those about its functions. Why are people so motivated to explain, and what accounts for our systematic explanatory preferences? While many plausible functions for explanation have been proposed, both philosophers and psychologists have emphasized that explanations could be valuable because they scaffold the kind of learning that supports adaptive behavior. For example, Craik (1943) described explanation as “a kind of distance-receptor in time, which enables organisms to adapt themselves to situations that are about to arise.” Heider (1958) suggested that we explain events in order to relate them to more general processes, allowing us “to attain a stable environment and have the possibility of controlling it.” In other words, explanations put us in a better position to predict and control the future.

Gopnik (2000) provocatively compares explanation to orgasm, suggesting that the phenomenological satisfaction of explanation is our evolutionarily provided incentive to engage in theory formation, as orgasm is to reproduction. She thus explains a characteristic of the process of explanation—its typical phenomenological outcome—by appeal to the function of its products—namely useful theories about the world. But the very process of engaging in explanation could have additional benefits. For example, attempting but failing to produce accurate explanations could nonetheless support future learning (e.g., Needham & Begg, 1991) and guide future inquiry (e.g., Legare, in press), and the effects of explanation could differ depending on whether the explanations are self-generated (involving process plus product) or provided (involving only product; see, e.g., Brown & Kane, 1988; Crowley & Siegler, 1999; Rittle-Johnson, 2006 for relevant discussion and findings).

The hypothesis that explanation supports adaptive behavior predicts that seeking, generating, and evaluating explanations should have systematic cognitive

consequences. The rest of the chapter is principally devoted to documenting these consequences. Before moving on, however, a few caveats are in order. First, explanations need not share a single, common function. Accordingly, the section that follows considers different kinds of explanations and their respective consequences for cognition. And while the present focus is on the roles of explanation in learning and inference, it is worth acknowledging that explanation certainly serves additional functions, as in persuasion and the assignment of blame.

A second caveat is that whatever the functions of explanation, it is unlikely that explainers have the explicit goal of fulfilling them. The basis for explanatory judgments could be opaque to introspection or based on indirect cues, no matter that their ultimate function is to achieve particular goals. And finally, it is quite likely that many of the properties of explanation stem not from any particular benefit (be it the result of natural selection or learning) but rather arise as a side effect of other cognitive characteristics. For example, a preference for simpler explanations could be a side effect of limited working memory. These caveats, however, do not detract from the value of examining the role of explanation in learning and inference.

Different Kinds of Explanations

Not all explanations will necessarily have a common structure or function. In fact, there are a variety of proposals suggesting fundamentally different kinds of explanations. For example, within philosophy, explanations for particular events or properties are often distinguished from explanations for laws or regularities, as are “how possibly” explanations—which explain how something *could* have come about—from “how actually” explanations—which explain how something *in fact* did come about (e.g., Salmon, 1989). Within psychology, advocates for domain specificity have suggested that different domains involve unique explanatory schemata (e.g., Carey, 1985; Wellman & Gelman, 1992). And explanations for why something is the case can be distinguished from explanations for why someone believes something to be the case (e.g., Kuhn, 2001). Depending on the particular account of explanation in question, these distinctions can correspond to explanations that differ in structure, in function, in both, or in neither.

A taxonomy that has proven particularly fruitful in psychology, with roots in Aristotle (see also Dennett, 1987), identifies (at least) three kinds

of explanations: mechanistic explanations, which explain by appeal to parts and processes; teleological or functional explanations, which cite functions or goals; and formal explanations, which cite kind or category membership. For example, consider explanations for why a particular tire is round. Explaining the roundness by appeal to the tire’s manufacturing process would qualify as mechanistic; explaining by appeal to its function in generating efficient movement would qualify as functional; and explaining by appeal to the objects’ category membership (“it’s round because it’s a tire”) would qualify as formal (see Lombrozo & Carey, 2006, for an extended discussion of mechanistic and functional explanations; see Prasada & Dillingham, 2006, for a discussion of formal explanations).

A growing body of evidence suggests that these three kinds of explanation correspond to distinct cognitive processes and representations. Keil (1992, 1994, 1995), for example, posits distinct explanatory stances or “modes of construal” corresponding to mechanistic and functional explanations, and finds that children adopt explanatory stances preferentially in different domains. Developing earlier suggestions by Sully (1900) and Piaget (1929), Kelemen (1999) suggests that functional explanations are an explanatory “default” that comes more naturally to children than do mechanistic explanations. It appears that children in fact do prefer such explanations “promiscuously,” as do adults who have not been exposed to alternative scientific explanations (Casler & Kelemen, 2008), who are required to respond under speeded conditions (Kelemen & Rosset, 2009), or who suffer from Alzheimer’s disease (Lombrozo, Kelemen, & Zaitchik, 2007).

Although functional explanations appeal to effects rather than to causal mechanisms, they are not necessarily inconsistent with causal theories of explanations. In fact, Lombrozo and Carey (2006) find evidence that functional explanations are only accepted when the function invoked in the explanation played a causal role in bringing about what is being explained (or more precisely, when the function is a token of a type that played a causal role in bringing about what is being explained; see also Wright, 1976, for analysis from philosophy). Thus, a tiger’s stripes can be explained by appeal to camouflage because the appeal to camouflage is shorthand for a complex causal process, whereby the stripes of past tigers supported camouflage, and this in turn played a causal role in the existence of the current, striped tiger.

Mechanistic and functional explanations can thus both be construed as causal explanations, but there is reason to think they reflect differences in underlying reasoning. First, while functional explanations ground out in a causal story, that story can be unboundedly complex and largely unknown. Most people are happy to explain a tiger's stripes by appeal to camouflage without any knowledge of tigers' evolutionary history, and with almost no knowledge of how natural selection operates (e.g., Shtulman, 2006). This suggests that information about causal mechanisms plays a different role for functional explanations than for mechanistic explanations. Second, mechanistic and functional explanations highlight different information. Mechanistic explanations privilege causal mechanisms and underlying constituents, while functional explanations privilege functions, intentions, goals, and design.

Consistent with this observation, Lombrozo (2010b) finds that the criteria for causal ascription differ depending on whether a causal system is construed functionally or mechanistically: When a system is construed mechanistically, causal ascriptions are more sensitive to the mechanism of transmission by which a cause brought about an effect. For example, a causal factor that contributed to an outcome indirectly, by preventing another factor from preventing the outcome, will receive a lower causal rating from a "mechanistic stance," but not necessarily from a "functional stance." Additional findings suggest that functional and mechanistic explanations have differential effects on categorization (Lombrozo, 2009) and generalization (Lombrozo & Gwynne, unpublished data), and are discounted asymmetrically (Heussen, 2010).

The unique explanatory contributions of mechanistic and functional explanations may be familiar to psychologists acquainted with Marr's levels of analysis (Marr, 1982). Marr proposed that psychological phenomena can be explained at three levels, including the *computational*, which involves a specification of the goals of a system; the *algorithmic*, which identifies the representations and processes involved in carrying out the specified goal; and the *implementational*, which characterizes how the representations and processes are physically instantiated. In specifying the goals of a cognitive system, a computational-level analysis typically supports functional explanations, while the implementation level analysis typically supports mechanistic explanations. The algorithmic level can support both kinds of explanations. For example, representations

of binocular disparity can be explained (functionally) by appeal to their role in computing depth information, and also (mechanistically) by appeal to optical, retinal, and cortical processes.

The final category of "Aristotelian" explanation that has received empirical attention is formal explanation, the study of which has been pioneered by Prasada and colleagues (Prasada & Dillingham, 2006, 2009). They find that only some properties can be explained by appeal to category membership. For example, participants are willing to explain a tire's roundness by appeal to its category membership ("it's round because it's a tire"), but much less willing to thus explain a tire's blackness ("it's black because it's a tire"), even though both share the same statistical association with tires. Properties that support formal explanations have a variety of related characteristics, such as licensing normative claims (e.g., "tires ought to be round").

In sum, explanations can be classified in a number of ways. The distinction between mechanistic, functional, and formal explanations has proven particularly fruitful, as each type of explanation appears to highlight different information and have unique cognitive consequences. This distinction could also underlie differences in the types of explanations typically observed across domains (for discussion, see Lombrozo & Carey, 2006). However, there are likely to be a number of alternative classifications with potential implications for our understanding of both the structure and functions of explanations. For example, the discussion so far has focused on how *selected* explanations—the *products*—can be productively classified, but complete explanations and explanatory processes could be similarly varied. Documenting and understanding the heterogeneity of explanation is thus an important focus for future research.

Explanation and Learning

Given the intimate relationship between explanation and understanding, it is no surprise that explanation has a profound impact in learning. There are at least three ways in which explanation can influence learning. First, there is the matter of which explanations are sought, which constrains what one learns about the environment. For example, upon first encountering an elephant you're likely to wonder why it has a trunk, but less likely to wonder why the number of its legs is a perfect square. Second, processes involved in the evaluation of explanations can influence what is learned

from provided explanations, be it in educational or everyday situations. And third, the very process of generating explanations, be it for oneself or others, can influence one's own understanding and ability to generalize to novel contexts. These three impacts on learning are considered in turn.

Although explanation seeking is pervasive, people are highly selective in what they seek explanations about. This selectivity is driven in large part by what the learner already knows. For example, it seems unlikely that someone will seek an explanation for why fire trucks are red if that person already knows (or believes she knows) the answer. This is not surprising if explanation seeking is a mechanism for learning. However, generating questions requires knowing enough to appreciate what one doesn't know (Miyake & Norman, 1979). Questions are thus highly constrained by prior beliefs. For example, we ask questions for which we expect sensible answers—even young children are more likely to ask what artifacts “are for” than what animals “are for” (Greif, Kemler-Nelson, Keil, & Guitierrez, 2006).

One attractive hypothesis is that people ask questions about events or properties that violate their expectations (Isaacs, 1930; Sully, 1900; see also Berlyne, 1954; Chouinard, 2007) or concern abnormal events (Hilton, 1990; Hilton & Slugoski, 1986)—a natural prediction if explanation is to guide learners beyond what they already know. Consistent with this proposal, analyses of why-questions reveal a high prevalence of negation (e.g., “Why *didn't* X occur?”), suggesting that the questioner is grappling with an unexpected observation (Hood & Bloom, 1979; but see Callanan & Oakes, 1992). And in experimental contexts, children are more likely to explain why a block did or did not make a machine light up when the observed outcome is inconsistent with previous training (Legare, Gelman, & Wellman, 2009). However, it is quite difficult to specify when an event counts as unexpected or surprising—it surely *isn't* the case that observations with low prior probability always prompt explanation, as most observations are in fact highly unlikely (consider, for example, any long sequence of coin flips: heads, heads, tails, heads, tails...).

Questions are also, of course, guided by interests: Children seem most likely to ask about the social world (Hickling & Wellman, 2001; Hood & Bloom, 1979), although questions about the physical world are not uncommon (Callanan & Oakes, 1992). There is also some evidence that from preschool to high school, children become increasingly likely to

ask questions motivated by potential applications rather than by general curiosity (Baram-Tsabari & Yarden, 2005). Finally, questions are guided by the goals of the learner. Students aiming to learn about musical instruments in order to design a novel instrument, for example, ask different questions from those exposed to the same material with the goal of identifying a promising musical group (Graesser, Langston, & Baggett, 1993). Providing a general characterization of what, when, and why people ask is complicated by the fact that interests and question-asking are highly contingent on both the explanation seeker and the context.

Once an explanation has been offered, what are the consequences for the recipient? First, explanations are often accompanied by a sense of understanding. For example, both undergraduates and clinicians who could explain an individual's symptoms perceived those symptoms to be more “normal” (Ahn, Novick, & Kim, 2003). And in some cases, provided explanations can be highly beneficial. As one example, children's performance on false belief tasks is related to how often their mothers explain mental states (Peterson & Slaughter, 2003). Unfortunately, however, the *sense* of understanding fostered by explanations is often an unreliable guide to actual understanding (Trout, 2002, 2008), and merely receiving an explanation is often insufficient to generate actual understanding. In particular, explanations provided in instructional contexts are frequently ineffective (Wittwer & Renkl, 2008). Explanations are more likely to support learning if they are appropriately tailored to what the learner already knows and appeal to general concepts or principles (see Wittwer & Renkl, 2008, for review). Explanations can also be more effective when they occur in the context of an interactive cognitive activity (Chi, 2009), in part because the recipient is able to request additional information and receive feedback. For example, young children are more likely to repeat questions after receiving a nonexplanation or an explanation they deem inadequate (Frazier, Gelman, & Wellman, 2009; see also Chouinard, 2007), and adult learners will request further elaboration when experts provide explanations that underestimate their knowledge (Wittwer, Nuckles, & Renkl, 2008).

Perhaps surprisingly, *generating* explanations can be a more effective mechanism for learning than *receiving* explanation. This phenomenon has been demonstrated in the context of peer tutoring, where tutors often profit more than tutees (e.g., Hooper,

1992; Roscoe & Chi, 2008; Ross & Cousins, 1995). The learning benefit of engaging in explanation—be it to oneself or to others—is known as the self-explanation effect (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi, de Leeuw, Chiu, & LaVancher, 1994) and has been found for preschoolers through adults, for a range of educational materials, and for both declarative and procedural knowledge (for review, see Fonseca & Chi, 2010). In a typical experiment, one group of participants is prompted to explain to themselves as they study an expository text or worked examples, such as math problems. These participants are compared with those in one or more control groups who study the same material without a prompt to explain, often with an alternative task (e.g., thinking aloud) or matched for study time. The typical finding is that participants who explain outperform their nonexplaining peers on a posttest, with the greatest benefit for transfer problems that require going beyond the material presented.

Effects of self-explanation have also been demonstrated in complex domains that arguably require a conceptual change in development. For example, prompting children to explain the correct response on Piagetian number conservation tasks leads to greater improvement than prompting them to explain their own (potentially incorrect) response, or than merely receiving feedback on the correct response (Siegler, 1995). Similarly, prompting children to explain why a character searched for an item in a particular location in a false belief task leads to greater improvement than merely predicting the character's behavior and receiving feedback, with benefits that extend to other theory of mind tasks (Amsterlaw & Wellman, 2006). These findings attest to the power of explanation as an engine for learning.

A variety of plausible mechanisms underlying the effects of explanations on learning have been proposed, many of which could operate in concert. For example, generating explanations can help learners translate declarative knowledge into usable procedures (Chi et al., 1989, 1994), integrate new material with prior beliefs (Ahn, Brewer, & Mooney, 1992; Chi et al., 1994; Lombrozo, 2006), identify and repair gaps in knowledge (Chi, 2000), or resolve potential inconsistencies (Johnson-Laird, Girotto, & Legrenzi, 2004). Other researchers have suggested that prompts to explain can increase efforts to seek explanations, leading to deeper processing, greater engagement with the task, and the reinforcement

of successful strategies over unsuccessful alternatives (Siegler, 2002). More recent work in cognitive development suggests that explanations can sometimes scaffold causal learning by presenting children with an easier task than prediction, as explaining is supported by knowledge of the outcome or property being explained (Wellman & Liu, 2007). As evidence, successful explanations precede successful predictions by preschoolers reasoning about an agent's choice between a contaminated and an uncontaminated food (Legare, Wellman, & Gelman, 2009; see also Amsterlaw & Wellman, 2006).

In comparing participants prompted to explain with those who are not so prompted, much of the research on explanation and learning focuses on explanation as a process rather than a product. This approach sidesteps questions about what kinds of utterances count as explanations. Research that has coded participant utterances for the quantity and content of explanations has tended to adopt a very broad conception of explanation, for example, including most articulated inferences that go beyond the material provided (Chi et al., 1989, 1994). It can thus be difficult to relate aspects of the structure of explanations, such as those posited by theories of explanation from philosophy, to the functional consequences of explanation for learning.

A few more recent strands of research, however, aim to relate structural accounts of explanation to functional consequences, and in so doing to elucidate why explanation has the particular learning benefits observed. In particular, research has focused on two widely accepted features of explanations: explanations tend to relate the particular property or event being explained to broader principles or regularities, and explanations often invoke causal relationships and causal mechanisms (see Woodward, 2010 for review from philosophy; see Lombrozo, 2006 for review from psychology). These properties correspond to important features of subsumption/unification and causal theories of explanation from philosophy, respectively.

Illustrating the role of subsumption and unification, Williams and Lombrozo (2010) propose the subsumptive constraints account, according to which explanation can facilitate learning by encouraging learners to understand what they are trying to explain in terms of broad, unifying generalizations (see also Wellman & Liu, 2007). If explaining exerts this particular constraint on the kind of structure learners seek, engaging in explanation should drive learners to discover and explicitly represent such

generalizations. Williams and Lombrozo (2010) tested this prediction in several experiments involving category learning. Compared to control conditions in which participants described category members, thought aloud during study, or engaged in free study, participants who explained were more likely to discover subtle, broad regularities underlying the category structure. Moreover, because the subsumptive constraints account predicts a selective advantage for explanation in discovering patterns, and not an “all-purpose” benefit in learning, it makes the novel prediction that explaining can *impair* learning when the material being explained does not support explanatory patterns, and recent findings confirm this prediction (Williams, Lombrozo, & Rehder, 2010; Kuhn & Katz, 2009; see also Dawes, 1999).

Illustrating the role of causation and causal mechanisms in mediating the role of explanation in learning are recent findings from cognitive development. Legare, Wellman, and Gelman (2010) found that children’s explanations posited unobserved causal mechanisms, such as germs, that were not spontaneously considered as a basis for prediction. This result suggests that explanations direct children to aspects of causal structure that might not be taken into account in the absence of explanatory processes (see also Legare, Gelman, & Wellman, 2009). Children who are prompted to explain how a novel, mechanical toy works also outperform peers who are prompted to observe the toy for a matched amount of time on measures that assess an understanding of causal mechanisms, but not on measures that involve memory for mechanism-irrelevant details, such as the toy’s colors (Legare & Lombrozo, unpublished data). Again, the findings suggest that explanation is not an all-purpose strategy for learning, but rather a highly selective process that influences precisely what a learner discovers and represents.

A great deal remains to be learned about the mechanisms by which explanation influences learning and how distinct mechanisms interact. Characterizing these mechanisms is important not only for understanding explanation but also for understanding the very nature of human learning and representation. One reason the self-explanation effect is so intriguing is because it challenges a simple picture of learning, modeled after learning from observations or testimony, according to which learning is identified with the acquisition of new information from the external world. Learning by explaining to oneself or to others—much like thought experiments—reveals a kind of learning that involves the reorganization,

rerepresentation, or inferential consequences of what is already known. Understanding this kind of learning could well provide a better template for understanding all types of learning (for relevant discussions of learning see Rips et al., Chapter 11; Buehner & Cheng, Chapter 12; Holyoak, Chapter 13; Bassok & Novick, Chapter 21; and Koedinger & Roll, Chapter 40).

Explanation and Inference

Although learning and inference are likely to involve similar mechanisms, research on explanation and inference has proceeded largely independently of research on explanation and learning. The former strand of research is motivated by the observation that everyday inferences face the problem of underdetermination: Many hypotheses are possible given what we know about the world. We generally don’t know for certain why the economy rallied, why the bus was late, whether a colleague’s excuse for a missed deadline was fact or fabrication, or whether Mrs. Peacock or Mr. Green committed the murder. Yet many decisions rest on identifying the best hypothesis or the probability of a given hypothesis. Explanation has thus been proposed as a mechanism for guiding such judgments, often referred to as “abductive inferences.”

The term “abduction” is associated with Charles Sanders Peirce, who distinguished abductive inferences, which posit explanatory hypotheses, from the broader class of inductive inferences, which also include inferences from a sample to a population, such as inferring that all swans are white after observing many white swans (Burch, 2010; see also Magnani, 2001). Subsequent scholars have not reliably distinguished abduction from induction, instead focusing on inferences to particular explanations. For example, Harman (1965) introduced the notion of “inference to the best explanation,” according to which “one infers, from the fact that a certain hypothesis would explain the evidence, to the truth of that hypothesis . . . one infers, from the premise that a given hypothesis would provide a ‘better’ explanation for the evidence than would any other hypothesis, to the conclusion that the given hypothesis is true” (p. 324). In other words, one relies on the existence or quality of an explanation to guide inference, an idea that has since been pursued both theoretically and empirically (see Lipton, 2004, for a nice treatment in philosophy).

Within psychology, there is mounting evidence that the quality of an explanation does serve as

a guide to its probability. Factors that boost an explanation's perceived quality include simplicity (Bonawitz & Lombrozo, in press; Lagnado, 1994; Lombrozo, 2007; Read & Marcus-Newhall, 1993; Thagard, 1989), breadth (Preston & Epley, 2005; Read & Marcus-Newhall, 1993; Thagard, 1989), consistency with prior knowledge (Thagard, 1989), and coherence (Pennington & Hastie, 1993). Several of these factors have also been shown to influence an explanation's perceived probability. For example, Lombrozo (2007) presented adults with novel diseases and symptoms. Participants learned about a patient with two symptoms that could be explained by appeal to a single disease (the "simple" explanation) or by appeal to the conjunction of two diseases (the "complex" explanation). In some conditions, participants were also presented with the base rates for the diseases, either in the form of summary frequencies or experienced cases. There were several notable results: (a) in the absence of base-rate information, participants overwhelmingly preferred the simple explanation; (b) when base-rate information was provided, the complex explanation had to be quite a bit more probable than the simpler explanation for a majority of participants to prefer it; (c) participants who selected the simple explanation when it was unlikely to be true overestimated the base-rate of the disease figuring in the simple explanation; yet (d) when the complex explanation was explicitly stated to be more likely, participants overwhelmingly chose it. These results suggest that in the face of probabilistic uncertainty, people treat simplicity as a probabilistic cue commensurate with base-rate information (see also Lu et al., 2008).

The strategy of relying on an explanation's quality as a guide to its probability appears to be widespread, but is it warranted? This has been a topic of lively debate within philosophy, and for some properties of explanations, such as simplicity, within statistics and computer science as well (see Baker, 2010, for review). For example, some have suggested that simpler explanations avoid "overfitting" data and hence will generalize more effectively to novel cases (e.g., Forster, 2000). Others have suggested that an initial bias toward simplicity results in more efficient learning insofar as a learner is likely to modify hypotheses fewer times (e.g., Kelly, 2004). Breadth and unification have more direct ties to probability, as explanations that are broader or more general will, by definition, apply to more cases and could also have stronger evidential support (Myrvold, 2003). However, it is also possible

that explanatory preferences result from cognitive limitations—for example, simpler explanations could be easier to process or remember—and that their influence on probability and inference is misguided. As one example, Pennington and Hastie (1992) find that presenting evidence to mock jurors in a chronological order, which makes the evidence easier to integrate into a coherent explanation for the events that occurred, systematically influences verdicts, in contrast to the predictions of most normative models.

In the cases considered so far, an explanation's quality is used as a guide to the probability of that *explanation*. Other findings suggest that the existence and quality of explanations can influence the judged probability of *what is being explained*. For example, Koehler (1991) reviews evidence that prompting participants to explain one outcome—say, why a Democrat might win the next election—boosts the probability assigned to that outcome. Similarly, explaining why a relationship holds—say, why people who are risk-averse might make better firefighters—increases the probability assigned to that relationship (e.g., Anderson & Sechler, 1986; see also Ross et al., 1977), as well as memory for the relevant correlation (Bower & Masling, unpublished data).

Thagard (1989) proposes a model of abductive inference that incorporates both of the influences just considered: how the quality of an explanation influences the degree of belief in that explanation *as well as* belief in what is explained. Thagard's approach, the Theory of Explanatory Coherence, involves seven principles that specify relationships between propositions and their consequences for belief. For example, the principles specify that propositions that participate in explanatory relationships (whether they do the explaining or are explained) are mutually reinforcing, as are those that provide analogous explanations. The model has been instantiated in a connectionist network and successfully used to model human judgments (e.g., Ranney & Thagard, 1988; Thagard, 2006), providing another source of evidence for the importance of explanatory evaluation in inference.

In addition to direct influences of explanatory evaluations on the probability assigned to an explanation and what it explains, explanations influence judgments in tasks that require assessing the probability of one claim in light of another. For example, Sloman (1994) provided participants with an initial claim, such as many secretaries "have a hard

time financing a house” or “have bad backs,” and asked them to evaluate the probability of a related claim, such as many furniture movers “have a hard time financing a house” or “have bad backs.” When the claims supported a common explanation (e.g., moderate income for trouble financing a house), participants provided higher probability estimates than when the claims supported different explanations (e.g., a sedentary job versus heavy lifting for back problems; see also Rehder, 2006). This could be because explanations compete (as might be expected in a framework like Thagard’s Explanatory Coherence), or because an explanation that applies to many cases is judged to be broader or more unifying and therefore of higher quality, resulting in a boost to the probability of what it explains (in keeping with the findings reviewed above).

The role of explanation in inference extends to a variety of additional judgments, including category membership (Rips et al., Chapter 11), decision making (LeBoeuf & Shafir, Chapter 16; see also Hastie & Pennington, 2000), and both legal (Spellman & Schauer, Chapter 36) and medical (Patel et al., Chapter 37) reasoning. For example, in the context of concept learning and categorization, explanatory knowledge has been shown to facilitate learning (Murphy & Allopenna, 1994; Williams & Lombrozo, 2010), influence judgments of the typicality of category members (Ahn, Marsh, Luhmann, & Lee, 2002; Murphy & Allopenna, 1994), and foster conceptual coherence (Murphy & Medin, 1985; Patalano, Chin-Parker, & Ross, 2006). In addition, the way in which a category’s features are explained can influence the relative importance of those features in making judgments about category membership: Features that support more explanations (Sloman, Love, & Ahn, 1998), causally deeper explanations (Ahn, 1998; Ahn & Kim, 2000), or explanations that are privileged by the explanatory stance adopted—mechanistic or functional (Lombrozo, 2009)—will typically be weighted more heavily. Such findings illustrate the breadth and importance of explanation’s effects.

Finally, although explanations likely benefit inference in many contexts, there are some well-documented cases in which explanatory judgments can lead people astray. For example, both children and adults tend to mistake explanations for *why* something is the case for evidence *that* something is the case (Glassner, Weinstock, & Neuman, 2005; Kuhn, 2001), especially when evidence is unavailable (Brem & Rips, 2000). People’s ability to detect

circularity in explanations is also imperfect (Baum, Danovitch, & Keil, 2008; Rips, 2002) and susceptible to extraneous factors, such as the presence of interesting but potentially irrelevant information (Weisberg, Keil, Goodstein, Rawson, & Gray, 2008). Finally, both children and adults tend to overestimate the accuracy and depth of their own explanations (Mills & Keil, 2004; Rozenblit & Keil, 2002). So while there is little doubt that explanation has a considerable influence on a wide range of everyday inferences, this influence is quite heterogeneous in scope and consequences, and almost certainly in underlying mechanisms.

Distinguishing Explanation From Other Cognitive Processes

Explanation is closely related to a variety of cognitive processes, including but not limited to inductive reasoning, deductive reasoning, categorization, causal reasoning, analogical reasoning, and learning. Can and should explanation be distinguished from these extant areas of study? From one perspective, the answer is clearly “no.” Explanation could trigger or be triggered by these processes, and it is likely to share many common mechanisms. Isolating explanation could thus result in a mischaracterization of cognitive mechanisms and their coordination. But there are also compelling reasons to study explanation and abductive inference as phenomena in their own right. The study of explanation highlights important aspects of learning, reasoning, and representation that are obscured when explanation is undifferentiated from general causal reasoning or inference. This section identifies a few distinctions that help locate the study of explanation and its unique contributions.

First, are explanation and abductive inference simply kinds of causal reasoning? One compelling reason to avoid this identification is because it assumes that all explanations are causal (see earlier section on “The Structure of Explanations”). There are broad classes of explanations that most people acknowledge as noncausal, such as mathematical and logical explanations. There are also explanations involving causal systems that are arguably noncausal. For example, consider an explanation for why a 1-inch square peg won’t fit into a 1-inch-diameter round hole. While causal factors are involved, it is arguably geometric facts that do the explanatory work (Putnam, 1975). The reverse also holds: Not all causal hypotheses are explanatory. It is the burden of a theory of explanation

to account for why the big bang is a poor explanation for cultural differences in color preferences, and more generally why some and only some causes strike us as explanations for their effects. Identifying “explanatory” causes—those that figure in a selected explanation—could be akin to what we do when we identify “the” cause of an effect or distinguish causes from enabling conditions, but it is certainly not equivalent to causal inference or to the specification of all causal factors.

Second, is it worth distinguishing explanations and abductive inference from the broader and quite heterogeneous class of inductive or deductive inferences? Again, not all explanations are straightforward examples of inference, and not all inferences are explanatory. Consider a case in which the causes of an event are firmly established: A match was struck in the presence of oxygen, and a fire ensued. In such a case, the striking of the match is likely to be selected as the explanation for the fire, not the presence of oxygen. This judgment requires an inference concerning which aspects of the causal structure are *explanatory*, but it need not be accompanied by an inference concerning which causes were present or type-level causal relationships, as these are already known. More important, explanatory inferences are a subset of everyday inferences. I can infer what the weather will be like based on a forecast, or what the 100th digit of pi is given the output of a computer program. But these inferences are not explanatory, and as a result the characteristics of abductive inference—such as a role for simplicity or breadth in determining the perceived probability of a hypothesis—need not apply.

Recognizing the unique aspects of explanation brings important questions into relief. For example, what is common between causal and noncausal explanations? Why do causal explanations privilege some kinds or aspects of causal structure over others? How do explanatory inferences go beyond inferences concerning causal and statistical structure? Do explanatory inferences involve unique mechanisms for evaluation? These questions can best be addressed through systematic investigation of the structure and functions of explanation, including new empirical findings to complement those reviewed here.

Conclusions

This chapter began by considering the role of explanation in cognition, and in particular why explanation is such a pervasive aspect of human

experience and elicits such strong intuitions. The reviewed findings confirm that the ability to evaluate and be guided by explanations is already quite sophisticated in young children and persists throughout adulthood. Moreover, explanatory preferences are highly systematic and have important consequences for core cognitive capacities that span learning and inference.

While explanations are quite heterogeneous and almost certainly impact cognition through a variety of distinct mechanisms, effects of explanations typically share a few characteristics. In particular, the process of explaining appears to recruit relevant prior beliefs along with a suite of formal constraints in the form of explanatory preferences—that is, preferences for explanations that have particular characteristics, such as simplicity or the ability to subsume and unify disparate observations. These properties of explanations (the products) account for many of the consequences of explanation (the process), which range from quite beneficial to neutral or even detrimental.

Despite a focus on both the functional consequences of explanations and how explanations are evaluated, there have been few attempts to provide comprehensive computational- or algorithmic-level accounts of explanation. In the absence of computational-level accounts, explanations and explanatory inferences lack normative benchmarks against which to assess human performance. And in the absence of algorithmic-level accounts, the relationship between explanation and other cognitive capacities remains largely opaque. Developing such accounts should thus be a high priority as research on explanation and abductive inference moves forward. The final section of the chapter identifies additional, promising avenues for future research.

Future Directions

Research on explanation is impressive in its scope and diversity, spanning cognitive psychology, social psychology, developmental psychology, education, philosophy, and beyond. However, distinct strands of research have proceeded almost independently, and until recently much of this work has not been unified under a common umbrella. A key direction for future research, then, involves the integration of different perspectives and findings across these areas. However, existing research raises more questions than it answers, and many aspects of explanations are only beginning to be explored. This section highlights a few questions that are ripe for further

investigation. The first six have already been raised in the course of the review; the final four involve additional issues.

1. Are there different kinds of explanations? If so, in virtue of what are they all explanatory?
2. What are the functions of explanation? To what extent are systematic explanatory preferences attributable to these functions, and to what extent do they result as side effects of other cognitive processes?
3. Which explanations are sought, and why? In particular, how do the characteristics of the explainer's prior beliefs interact with the environment to determine what is deemed to require explanation?
4. What are the mechanisms by which explanation guides learning? How is the role of explanation influenced by feedback on the accuracy of explanations or from further observations or testimony? When is the role of explanation beneficial, and when is it detrimental?
5. What are the criteria employed in the evaluation of explanations, and what are the consequences for learning and inference? In particular, what are the conditions under which explanatory considerations—such as an explanation's simplicity—inform assessments of probability, and is this role for explanatory considerations ever warranted?
6. What are the prospects for a normative or computational-level theory of explanation and abductive inference? How will such a theory inform empirical research, and how should the theory be constrained by empirical findings?
7. How do social and pragmatic factors influence explanations and their consequences?
8. To what extent are there individual differences and cultural variation in explanatory preferences? What is the source of this variation, and what are the consequences for learning and inference?
9. To what extent can explanation be separated from language? While typical explanations are encoded in language, they need not be. For example, Baillargeon (2004) proposed that prelinguistic infants construct explanations, while Povinelli and Dunphy-Lelii (2001) devised a clever method for assessing “explanation” in chimpanzees. As a conceptual question, how can explanation be characterized without a commitment to some language-like manifestation? And as a methodological

question, how can nonlinguistic explanations be studied empirically?

10. How are explanations generated? There has been much more research on how explanations are evaluated than on how they are generated. This is in part because explanation generation confronts many of the most difficult questions in cognitive science concerning the content and structure of general beliefs and how these are retrieved in particular contexts. How can the study of explanation generation be informed by research on representation and memory, and how might findings concerning the generation of explanations in turn contribute to these areas?

Acknowledgments

Sincere thanks to Joseph Austerweil, Thomas Griffiths, Ulrike Hahn, Keith Holyoak, G. Randolph Mayes, and Joseph Williams for comments on earlier drafts of this chapter, and NSF grant DRL-1056712 for support..

References

- Ahn, W. (1998). Why are different features central for natural kinds and artifacts? *Cognition*, 69, 135–178.
- Ahn, W., & Bailenson, J. (1996). Causal attribution as a search for underlying mechanisms: An explanation of the conjunction fallacy and the discounting principle. *Cognitive Psychology*, 31, 82–123.
- Ahn, W., Brewer, W. F., & Mooney, R. J. (1992). Schema acquisition from a single example. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 391–412.
- Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation vs. mechanism information in causal attribution. *Cognition*, 54, 299–352.
- Ahn, W., & Kim, N. S. (2000). The causal status effect in categorization: An overview. In D. L. Medin (Ed.), *Psychology of learning and motivation* 40 (pp. 23–65) New York: Academic Press.
- Ahn, W., Marsh, J. K., Luhmann, C. C., & Lee, K. (2002). Effect of theory-based feature correlations on typicality judgments. *Memory and Cognition*, 30, 107–118.
- Ahn, W., Novick, L., & Kim, N. S. (2003). Understanding it makes it more normal. *Psychonomic Bulletin and Review*, 10, 746–752.
- Amsterlaw, J., & Wellman, H. (2006). Theories of mind in transition: A microgenetic study of the development of false belief understanding. *Journal of Cognition and Development*, 7, 139–172.
- Anderson, C. A., Krull, D. S., & Weiner, B. (1996). Explanations: Processes and consequences. In E.T. Higgins & A.W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 271–296). New York: Guilford Press.
- Anderson, C. A., & Sechler, E. S. (1986). Effects of explanation and counterexplanation on the development and use of social theories. *Journal of Personality and Social Psychology*, 50, 24–34.
- Baker, A. (2010). Simplicity. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2010 ed.). URL Retrieved August 2011, from <http://plato.stanford.edu/archives/spr2010/entries/simplicity/>

- Baillargeon, R. (2004). Infants' physical world. *Current Directions in Psychological Science*, 13, 89–94.
- Baram-Tsabari, A., & Yarden, A. (2005). Characterizing children's spontaneous interests in science and technology. *International Journal of Science Education*, 27, 803–826.
- Baum, L. A., Danovitch, J. H., & Keil, F. C. (2008). Children's sensitivity to circular explanations. *Journal of Experimental Child Psychology*, 100, 146–155.
- Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. London: Routledge.
- Berlyne, D. E. (1954). A theory of human curiosity. *British Journal of Psychology*, 45, 180–191.
- Bonawitz, E.B. & Lombrozo, T. (in press). Occam's rattle: children's use of simplicity and probability to constrain inference. *Developmental Psychology*.
- Brem, S. K., & Rips, L. J. (2000) Explanation and evidence in informal argument. *Cognitive Science*, 24, 573–604.
- Brown, A. L., & Kane, M. J. (1988). Preschool children can learn to transfer: Learning to learn and learning from example. *Cognitive Psychology*, 20, 493–523.
- Burch, R. (2010). Charles Sanders Peirce. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2010 ed.). Retrieved August 2011, from <http://plato.stanford.edu/archives/fall2010/entries/peirce/>
- Callanan, M. A., & Oakes, L. (1992). Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. *Cognitive Development*, 7, 213–233.
- Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.
- Casler, K., & Kelemen, D. (2008). Developmental continuity in the teleo-functional explanation: Reasoning about nature among Romanian Romani adults. *Journal of Cognition and Development*, 9, 340–362.
- Chi, M. T. H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 161–238). Hillsdale, NJ: Erlbaum.
- Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, 1, 73–105.
- Chi, M. T. H., Bassok, M., Lewis, M., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13, 145–182.
- Chi, M. T. H., de Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18, 439–477.
- Chin-Parker, S., & Bradner, A. (2010). Background shifts affect explanatory style: How a pragmatic theory of explanation accounts for background effects in the generation of explanations. *Cognitive Processing*, 11, 227–249.
- Chouinard, M. (2007). Children's questions: A mechanism for cognitive development. *Monographs of the Society for Research in Child Development*, 72, 1–57.
- Craik, K. (1943). *The nature of explanations*. Cambridge, England: Cambridge University Press.
- Craver, C. (2007). *Explaining the brain: What a science of the mind-brain could be*. New York: Oxford University Press.
- Crowley, K., & Siegler, R. S. (1999). Explanation and generalization in young children's strategy learning. *Child Development*, 70, 304–316.
- Darden, L. (2006). *Reasoning in biological discoveries: Essays on mechanisms, interfiled relations, and anomaly resolution*. Cambridge, England: Cambridge University Press.
- Dawes, R.M. (1999). A message from psychologists to economists: Mere predictability doesn't matter like it should (without a god story appended to it). *Journal of Economic Behavior & Organization*, 39, 29–40.
- DeJong, G. (2004). Explanation-based learning. In A. Tucker (Ed.), *Computer science handbook* (2nd ed., pp. 68–1–68–20).
- Dennett, D. (1987). *The intentional stance*. Cambridge, MA: MIT Press.
- Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, 99, 3–19.
- Fonseca, B. A., & Chi, M. T. H. (2010). Instruction based on self-explanation. In R. Mayer & P. Alexander (Eds.), *The handbook of research on learning and instruction* (pp. 296–321). New York: Routledge Press.
- Forster, M. R. (2000). Key concepts in model selection – performance and generalizability. *Journal of Mathematical Psychology*, 44, 205–231.
- Frazier, B. N., Gelman, S. A., & Wellman, H. M. (2009). Preschoolers' search for explanatory information within adult-child conversation. *Child Development*, 80, 1592–1611.
- Friedman, M. (1974). Explanation and scientific understanding. *Journal of Philosophy*, 71, 5–19.
- Garfinkel, A. (1990). *Forms of explanation: Rethinking the questions in social theory*. New Haven, CT: Yale University Press.
- Glassner, A., Weinstock, M., & Neuman, Y. (2005). Pupils' evaluation and generation of evidence and explanation in argumentation. *British Journal of Educational Psychology*, 75, 105–118.
- Goodman, N. D., Baker, C. L., Bonawitz, E. B., Mansinghka, V. K., Gopnik, A., Wellman, H., ..., Tenenbaum, J. B. (2006). Intuitive theories of mind: A rational approach to false belief. In R. Sun, N. Miyake, C. Schunn, & S. Lane (Eds.), *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society* (pp. 1382–1387).
- Gopnik, A. (2000). Explanation as orgasm and the drive for causal knowledge: The function, evolution, and phenomenology of the theory-formation system. In F. Keil & R. A. Wilson (Eds.), *Explanation and cognition* (pp. 299–324). Cambridge, MA: MIT Press.
- Gopnik, A., & Meltzoff, A. (1997). *Words, thoughts and theories*. Cambridge, MA: MIT Press.
- Graesser, A. C., Langston, M. C., & Baggett, W. B. (1993). Exploring information about concepts by asking questions. In G. V. Nakamura, R. M. Taraban, & D. Medin (Eds.), *The psychology of learning and motivation. Vol. 29: Categorization by humans and machines* (pp. 411–436). Orlando, FL: Academic Press.
- Greif, M., Kemler-Nelson, D., Keil, F. C., & Gutierrez, F. (2006). What do children want to know about animals and artifacts? Domain-specific requests for information. *Psychological Science*, 17, 455–459.
- Harman, G. (1965). The inference to the best explanation. *Philosophical Review*, 74, 88–95.
- Hart, H. L. A., & Honnoré, T. (1985). *Causation in the law* (2nd ed.). New York: Oxford University Press.
- Hastie, R., & Pennington, N. (2000) Explanation-based decision making. In T. Connolly, H. R. Arkes, & K. R. Hammond (Eds.), *Judgment and decision making: An interdisciplinary reader* (2nd ed., pp. 212–228). Cambridge, England: Cambridge University Press.

- Heider, F. (1958). *The psychology of interpersonal relations*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hempel, C. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. New York: Free Press.
- Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, 15, 135–175.
- Heussen, D. S. (2010). When functions and causes compete. *Thinking and Reasoning*, 16, 233–250.
- Hickling, A. K., & Wellman, H. M. (2001). The emergence of children's causal explanations and theories: Evidence from everyday conversation. *Developmental Psychology*, 37, 668–683.
- Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, 107, 65–81.
- Hilton, D. J. (1996). Mental models and causal explanation: Judgments of probable cause and explanatory relevance. *Thinking and Reasoning*, 2, 273–308.
- Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, 93, 75–88.
- Hood, L., & Bloom, L. (1979). What, when, and how about why: A longitudinal study of the early expressions of causality. *Monographs of the Society for Research in Child Development*, 44(6, serial no. 181).
- Hooper, S. (1992). Effects of peer interaction during computer-based mathematics instruction. *Journal of Educational Research*, 85, 180–189.
- Isaacs, N. (1930). Children's "why" questions. In S. Isaacs (Ed.), *Intellectual growth in young children* (pp. 291–349). London: Routledge.
- Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review*, 111, 640–661.
- Keil, F. C. (1992). The origins of an autonomous biology. In M. R. Gunnar & M. Maratsos (Eds.), *Modularity and constraints in language and cognition. Vol. 25: Minnesota Symposium on Child Psychology* (pp. 103–138). Hillsdale, NJ: Erlbaum.
- Keil, F. C. (1994). The birth and nurturance concepts by domains: The origins of concepts of living things. In L. A. Hirschfeld & S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 234–254). Cambridge, England: Cambridge University Press.
- Keil, F. C. (1995). The growth of causal understanding of natural kinds. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multi-disciplinary debate* (pp. 234–262). Oxford, England: Clarendon Press.
- Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Science*, 7, 368–373.
- Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, 57, 227–254.
- Keil, F. C., & Wilson, R. A. (2000). *Explanation and cognition*. Cambridge, MA: The MIT Press.
- Kelemen, D. (1999) Function, goals and intention: Children's teleological reasoning about objects. *Trends in Cognitive Science*, 3, 461–468.
- Kelemen, D., & Rosset, E. (2009). The human function compunction: Teleological explanation in adults. *Cognition*, 111, 138–143.
- Kelly, K. (2004). Justification as truth-finding efficiency: How Ockham's Razor works. *Minds and Machines*, 14, 485–505.
- Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. Salmon (Eds.), *Scientific explanation* (pp. 410–505). Minneapolis: University of Minnesota Press.
- Koehler, D. J. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin*, 110, 499–519.
- Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. Cambridge, MA: MIT Press.
- Krull, D. S., & Anderson, C. A. (2001). Explanation, cognitive psychology of. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (Vol. 8, pp. 5150–5154). Oxford, England: Elsevier.
- Kuhn, D. (2001). How do people know? *Psychological Science*, 12, 1–8.
- Kuhn, D., & Katz, J. (2009). Are self-explanations always beneficial? *Journal of Experimental Child Psychology*, 103, 386–394.
- Lagnado, D. (1994). *The psychology of explanation: A Bayesian approach*. Unpublished Masters thesis. Schools of Psychology and Computer Science, University of Birmingham, England.
- Legare, C. H. (in press). Exploring explanation: Explaining inconsistent evidence informs exploratory hypothesis-testing behavior in young children. *Child Development*.
- Legare, C. H., Gelman, S. A., & Wellman, H. M. (2010). Inconsistency with prior knowledge triggers children's causal explanatory reasoning. *Child Development*, 81, 929–944.
- Legare, C. H., Wellman, H. M., & Gelman, S. A. (2009). Evidence for an explanation advantage in naïve biological reasoning. *Cognitive Psychology*, 58, 177–194.
- Lipton, P. (2004). *Inference to the best explanation*. New York: Routledge.
- Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, 10, 464–470.
- Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, 55, 232–257.
- Lombrozo, T. (2009). Explanation and categorization: How "why?" informs "what?" *Cognition*, 110, 248–253.
- Lombrozo, T. (2010a). From conceptual representations to explanatory relations. *Behavioral and Brain Sciences*, 33, 218–219.
- Lombrozo, T. (2010b). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, 61, 303–332.
- Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, 99, 167–204.
- Lombrozo, T., Kelemen, D., & Zaitchik, D. (2007). Inferring design: Evidence of a preference for teleological explanations in patients with Alzheimer's Disease. *Psychological Science*, 18, 999–1006.
- Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, 115, 955–982.
- Machery, E. (2009). *Doing without concepts*. New York: Oxford University Press.
- Mackie, J. L. (1980). *The cement of the universe: A study of causation*. New York: Oxford University Press.
- Magnani, L. (2001). *Abduction, reason, and science. Processes of discovery and explanation*. New York: Kluwer Academic/Plenum Publishers.
- Malle, B. F. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Cambridge, MA: MIT Press.
- Margolis, E. (1995). The significance of the theory analogy in the psychological study of concepts. *Mind and Language*, 10, 45–71.
- Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. New York: W. H. Freeman.

- Mills, C., & Keil, F. C. (2004). Knowing the limits of one's understanding: The development of an awareness of an illusion of explanatory depth. *Journal of Experimental Child Psychology*, 87, 1–32.
- Miyake, N., & Norman, D. A. (1979). To ask a question one must know enough to know what is not known. *Journal of Verbal Learning and Verbal Behavior*, 18, 357–364.
- Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 20, 904–919.
- Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92, 289–316.
- Myrvold, W. C. (2003). A Bayesian account of the virtue of unification. *Philosophy of Science*, 70, 399–423.
- Needham, D. R., & Begg, I. M. (1991). Problem-oriented training promotes spontaneous analogical transfer: Memory-oriented training promotes memory for training. *Memory and Cognition*, 19, 543–557.
- Patalano, A. L., Chin-Parker, S., & Ross, B. H. (2006). The importance of being coherent: Category coherence, cross-classification, and reasoning. *Journal of Memory and Language*, 54, 407–424.
- Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the story-model for juror decision making. *Journal of Personality and Social Psychology*, 62, 189–206.
- Pennington, N., & Hastie, R. (1993). The story model for juror decision making. In R. Hastie (Ed.), *Inside the juror: The psychology of juror decision making* (pp. 192–221). Cambridge, England and New York: Cambridge University Press.
- Peterson, C., & Slaughter, V. (2003). Opening windows into the mind: Mothers' preferences for mental state explanations and children's theory of mind. *Cognitive Development*, 18, 399–429.
- Piaget, J. (1929). *The child's conception of the world*. London: Routledge & Kegan Paul.
- Povinelli, D. J., & Dunphy-Lelii, S. (2001). Do chimpanzees seek explanations? Preliminary comparative investigation. *Canadian Journal of Experimental Psychology*, 55, 185–193.
- Prasada, S., & Dillingham, E. M. (2006). Principled and statistical connections in common sense conception. *Cognition*, 99, 73–112.
- Prasada, S., & Dillingham, E. M. (2009). Representation of principled connections: A window onto the formal aspect of common sense conception. *Cognitive Science*, 33, 401–448.
- Preston, J., & Epley, N. (2005). Explanations versus applications: The explanatory power of valuable beliefs. *Psychological Science*, 16, 826–832.
- Putnam, H. (Ed.) (1975). Philosophy and our mental life. In *Mind, language and reality: Philosophical papers* (Vol. 2, pp. 291–303). Cambridge, England: Cambridge University Press.
- Railton, P. (1981). Probability, explanation, and information. *Synthese*, 48, 233–256.
- Ranney, M., & Thagard, P. (1988). Explanatory coherence and belief revision in naive physics. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 426–432). Hillsdale, NJ: Erlbaum.
- Read, S. J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, 65, 429–447.
- Rehder, B. (2006). When causality and similarity compete in category-based property induction. *Memory and Cognition*, 34, 3–16.
- Rips, L. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21–59). Cambridge, England: Cambridge University Press.
- Rips, L. J. (2002). Circular reasoning. *Cognitive Science*, 26, 767–795.
- Rittle-Johnson, B. (2006). Promoting transfer: The effects of direct instruction and self-explanation. *Child Development*, 77, 1–15.
- Roscoe, R. D., & Chi, M. T. H. (2008). Tutor learning: The role of explaining and responding to questions. *Instructional Science*, 36(4), 321–350.
- Ross, J., & Cousins, J. B. (1995). Giving and receiving explanations in cooperative learning groups. *Alberta Journal of Educational Research*, 41, 104–122.
- Ross, L., Lepper, M. R., Strack, F., & Steinmetz, J. (1977). Social explanation and social expectation: Effects of real and hypothetical explanations on subjective likelihood. *Journal of Personality and Social Psychology*, 35, 817–829.
- Rozenblit, L. R., & Keil, F. C. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, 26, 521–562.
- Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.
- Salmon, W. (1989). *Four decades of scientific explanation*. Minneapolis: University of Minnesota Press.
- Shtulman, A. (2006). Qualitative differences between naive and scientific theories of evolution. *Cognitive Psychology*, 52, 170–194.
- Shultz, T. R. (1980). Rules of causal attribution. *Monographs of the Society for Research in Child Development*, 47, 1–51.
- Siegler, R. S. (1995). How does change occur: A microgenetic study of number conservation. *Cognitive Psychology*, 28, 225–273.
- Siegler, R. S. (2002). Microgenetic studies of self-explanations. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 31–58). New York: Cambridge University Press.
- Sloman, S. A. (1994). When explanations compete: The role of explanatory coherence on judgments of likelihood. *Cognition*, 52, 1–21.
- Sloman, S. A., Love, B. C., & Ahn, W. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, 22, 189–228.
- Strevens, M. (2008). *Depth: An account of scientific explanation*. Cambridge, MA: Harvard University Press.
- Sully, J. (1900). *Studies of childhood*. New York: D. Appleton & Company.
- Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, 12, 435–467.
- Thagard, P. (2000). Probabilistic networks and explanatory coherence. *Cognitive Science Quarterly*, 1, 93–116.
- Thagard, P. (2006). Evaluating explanations in science, law, and everyday life. *Current Directions in Psychological Science*, 15, 141–145.
- Trout, J. D. (2002). Scientific explanation and the sense of understanding. *Philosophy of Science*, 69, 212–233.
- Trout, J. D. (2008). Seduction without cause: Uncovering explanatory neurophilia. *Trends in Cognitive Sciences*, 12, 281–282.
- van Fraassen, B. C. (1980). *The scientific image*. Oxford, England: Oxford University Press.
- Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience*, 20(3), 470–477.
- Wellman, H. M. (2011). Reinvigorating explanations for the study of early cognitive development. *Child Development Perspectives*, 5(1), 33–38.

- Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, 43, 337–375.
- Wellman, H. M., & Liu, D. (2007). Causal reasoning as informed by the early development of explanations. In A. Gopnik & L. E. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 261–279). New York: Oxford University Press.
- Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science*, 34, 776–806.
- Williams, J. J., Lombrozo, T., & Rehder, B. (2010). Why does explaining help learning? Insight from an explanation impairment effect. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2906–2911). Austin, TX: Cognitive Science Society.
- Wittwer, J., Nuckles, M., & Renkl, A. (2008). Is underestimation less detrimental than overestimations? The impact of experts' beliefs about a layperson's knowledge on learning and question asking. *Instructional Science*, 36, 27–52.
- Wittwer, J., & Renkl, A. (2008). Why instructional explanations often do not work: A framework for understanding the effectiveness of instructional explanations. *Educational Psychologist*, 43, 49–64.
- Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford, England: Oxford University Press.
- Woodward, J. (2010). Scientific explanation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2010 ed.). Retrieved August 2011, from <http://plato.stanford.edu/archives/spr2010/entries/scientific-explanation/>
- Wright, L. (1976). *Teleological explanations*. Berkeley: University of California Press.

Rational Argument

Ulrike Hahn and Mike Oaksford

Abstract

Argumentation is an integral part of how we negotiate life in a complex world. In many contexts it matters, furthermore, that arguments be rational, not that they are simply convincing. Rational debate is subject to both procedural norms and to epistemic norms that allow the evaluation of argument content. This chapter outlines normative frameworks for argumentation (dialectical, logical, and Bayesian); it then summarizes psychological research on argumentation, drawn from cognitive psychology, as well as a number of applied domains.

Key Words: argumentation, logic, probability, evidence, reasoning, inference, belief change

Introduction: The Realm of Argumentation

Argumentation is an integral part of many aspects of our complex social worlds: from law, politics, academia, and business, to our personal lives. Though the term “argument” often carries negative connotations in everyday life, many different types of argumentative dialog can be distinguished, such as quarrels, bargaining or negotiation, educational dialogues, and, central to the present chapter, critical discussion (see, e.g., van Eemeren & Grootendorst, 2004; Walton, 2006a). Argumentation in the sense of a critical discussion is about rational debate and has been defined as

...a verbal and social activity of reason aimed at increasing (or decreasing) the acceptability of a controversial standpoint for a listener or reader, by putting forward a constellation of propositions intended to justify (or refute) the standpoint before a “rational judge.” (van Eemeren, Grootendorst, & Snoeck Henkemans, 1996, p. 5)

It is in this sense of rational debate that psychological research on argumentation uses the

terms “argumentation” and “argument,” and it is the emphasis on rationality that distinguishes it from social psychological research on persuasion. Persuasion research has identified a wide range of factors that affect the degree to which a persuasive communication will be effective (see, e.g., Eagly & Chaiken, 1993; Maio & Haddock, 2010; also see Molden & Higgins, Chapter 20). In contrast to argumentation, persuasion research is concerned with “what works” and why, regardless of whether it is rational. Indeed, some of the aspects identified—such as mood, likeability of the speaker, or personal involvement of the listener—do not always lead to changes in convictions in ways that one might consider rational; nevertheless, many findings within the persuasion literature are also directly relevant to argumentation and will be discussed within this chapter. Within psychology, argumentation has also been the focus of developmental and education research. Here research has focused on the way children’s argumentation skills develop (e.g., Kuhn, 1989, 1991; Means & Voss, 1996) and examined ways in which critical thinking and argument skills

might be fostered (Kuhn, 1991, 2001). Finally, argumentation has close connections to the study of reasoning because inference as studied in reasoning research is an integral part of argument (on that relationship see, e.g., Chater & Oaksford, Chapter 2; Hahn & Oaksford, 2007a; Mercier & Sperber, 2011; Oaksford & Hahn, 2007).

Argumentation, however, has not only been of interest within psychology. Most of the extant literature on argumentation resides in philosophy, where the topic has been pursued since the Greek philosophers, in particular Aristotle (2004). Most directly, argumentation has been studied within philosophical logic; relevant research can also be found, however, within the philosophy of science and within epistemology (the philosophical study of knowledge and justified belief). In all of these contexts, philosophers have typically focused on normative theories, that is, theories of how we *should* behave.

Computer scientists, finally, have sought to devise novel frameworks for dealing with dialectical information, seeking to capture the structural relationships between theses, rebuttals, and supporting arguments with the degree of explicitness necessary for the design of computational argumentation systems. The uses to which argumentation has been put, in particular within artificial intelligence, are manifold (for an overview see, e.g., Bench-Capon & Dunne, 2007). For example, argumentation has been integral to attempts to build legal expert systems, that is, systems capable of providing support for legal advice (e.g., Prakken, 2008; also see Spellman & Schauer, Chapter 36). Argumentation-based frameworks have also been employed for medical decision support systems (e.g., Fox, Glasspool, Grecu, Modgil, South, & Patkar, 2007; see also Patel et al., Chapter 37) and have gained considerable attention in the context of multiagent systems (e.g., Rahwan & Moraitsis, 2009). Finally, computer scientists have developed software for the visualization of complex, interconnected sequences of arguments such as Araucaria (see e.g., Reed & Rowe, 2004). They have also developed software foundations for large-scale, socially contributed argumentative content on the Worldwide Web, which allow Web users to contribute, visualize, and analyze arguments on a particular theme (e.g., Rahwan, Zablith, & Reed, 2007).

In short, the wealth of practical contexts in which argumentation matters is matched by the interdisciplinary breadth with which the topic is studied. It should thus come as no surprise that there is also

more than one theoretical framework that can be brought to bear on the issue of good argument.

Theoretical Frameworks

Statements about rationality necessarily contain an evaluative component. Argument, like any behavior, is rational because it is “good” relative to some standard, whether this standard be norms seeking to guarantee truth or consistency, or a suitable functional relationship to the goals of the agent in a particular environment (e.g., Nickerson, 2008; Oaksford & Chater, 2007, 2009; see also Chater & Oaksford, Chapter 2; Griffiths et al., Chapter 3; Stanovich, Chapter 22).

Norms governing rational argument can crudely be classified into two broad categories: those aimed primarily at the content or structure of an argument, and those aimed at argumentative procedure. This distinction has its origins in two different philosophical projects aimed at characterizing what makes a good argument: so-called *epistemic* accounts, which are aimed at truth or justified belief, on the one hand, and so-called dialectical or *procedural* approaches, aimed at consensus, on the other.

For millennia, logic provided the normative framework for the evaluation of inference and, with that, argument: Logic provides an epistemic standard, enforcing consistency, and where reasoning proceeds from true premises, guaranteeing the truth of the conclusions that can be reached (see Evans, Chapter 8; Johnson-Laird, Chapter 9). However, logic is severely limited in its ability to deal with everyday informal argument (see also, e.g., Hamblin, 1970; Heyssse, 1997; Johnson, 2000; as well as Boger, 2005 for further references). In particular, logic seems poorly equipped to deal with the uncertainty inherent in everyday reason.

One of the first to bring these limitations to prominence was Toulmin (1958). According to Toulmin, an argument can be broken down into several distinct components. Essential to all arguments are a claim (i.e., a conclusion in need of support), evidence (i.e., facts that are used to support the claim), and warrants (i.e., the reasons that are used to justify the connections between the evidence and the claim). Warrants may receive further support through additional, auxiliary statements as “backing.” Any of these components may be subject to qualifiers such as “sometimes,” “frequently,” or “almost always,” and may be challenged by rebuttals, that is counterarguments, which themselves involve the same structural components. The identification

of these kinds of relationships has, for example, been integral to the visualization of argument structure and educational research.¹

In the context of argument, Toulmin was skeptical of absolute truth, and his model was inspired by the dialectical nature of the adversarial legal systems of the common law tradition, where a relative truth is established through a contest between proponents of two opposing positions. In the courtroom, the provision of evidence is constrained by procedural rules; for example, not all evidence is permissible. Hence, it is not the absolute truth about whether a defendant is guilty, for instance, that a trial seeks to establish, but a trial-relative truth, in the sense of what can be established within the constraints of those procedural norms.

Law has been taken as a leading example of argumentative dialog by many subsequent theorists (e.g., Rescher, 1977), and the argumentation literature is full of legally inspired concepts such as the notion of *burden of proof*. The most well-known burden of proof in law is the presumption of innocence in criminal trials: It is up to the prosecution to establish conclusively that a defendant is guilty, and failure to do so means the defendant goes free. Many argumentation theorists posit similar rules for rational discourse (see e.g., Walton, 1988; for a more critical perspective on the burden of proof in argumentation, see Hahn & Oaksford, 2007b).²

So-called pragma-dialectical theories of argumentation, in particular, have sought to identify the procedural norms governing rational debate (e.g., van Eemeren & Grootendorst, 1984, 1992, 2004; Walton, 1995, 1998a). The general flavor of these norms is well illustrated by a few rules from van Eemeren and Grootendorst's (2004) system:

Rule 2 states that

the discussant who has called the standpoint of the other discussant into question in the confrontation stage is always entitled to challenge the discussant to defend this standpoint. (p. 137)

The next rule, Rule 3, governs the burden of proof, whereby it is the proponent of a claim who is obliged to provide sufficient support for it:

The discussant who is challenged by the other discussant to defend the standpoint that he has put forward in the confrontation stage is always obliged to accept this challenge, unless the other discussant is not prepared to accept any shared premises and discussion rules; the discussant remains obliged to defend the standpoint as long as he does not retract it and as long as

he has not successfully defended it against the other discussant on the basis of the agreed premises and discussion rules. (p. 139)

In addition to such fundamental rules, theorists have proposed more specific rules governing, for instance, mode of presentation, and some examples of such rules will be discussed later.

The dialectical model of the courtroom, finally, has also been of influence in the development of nonclassical logics (e.g., Dung, 1995), which seek to overcome some of the inadequacies of classical logic, and, in particular, seek to capture reasoning under conditions of uncertainty (for a general overview of nonclassical logics, see Prakken & Vreeswijk, 2002).

Arguably, however, there is still something missing in procedural or dialectical approaches to argument as is readily apparent from the legal model itself: This is the “final evaluation.” In the courtroom, judges and/or juries reach a final verdict, and that overall evaluation itself should be rational. Procedural rules constrain this but do not fully determine its outcome. One may have several pieces of evidence, for example, all of which are permissible but which, individually, are more or less compelling. Furthermore, different pieces of evidence, potentially varying in individual strength, must be combined into an overall degree of support.³

Such evaluation is crucially about specific *content*. Procedural rules alone are insufficient here, but so typically are the purely structural relationships between statements identified by classical and nonclassical logical systems. Knowing, in structural terms, simply that one statement “attacks” another does not allow a final evaluation of whether one is more convincing than the other, and hence outweighs it in final evaluation.

Concerning content, it has most recently been argued that Bayesian probability might provide appropriate epistemic norms for argumentation (Hahn & Oaksford, 2006a,b 2007a; Korb, 2004; Oaksford & Hahn, 2004; also see Griffiths et al., Chapter 3).⁴ The Bayesian approach to argumentation originated as an attempt to provide a formal treatment of the traditional catalog of fallacies of argumentation, examples of which are circular arguments such as “God exists, because the Bible says so and the Bible is the word of God” and so-called arguments from ignorance, such as “Ghosts exist because nobody has proven that they don’t.” According to the Bayesian account, informal arguments such as these consist of a claim (“ghosts exist”) and evidence for that claim (“nobody has

proven that they don't"). An individual's degree of belief in the claim is represented by a probability. Bayes' theorem, which follows from the fundamental axioms of probability theory, then provides a normative standard for belief revision; it thus provides a formal tool for evaluating how convinced that individuals should be about the claim in light of that particular piece of evidence. There are three probabilistic quantities involved in Bayes' theorem that determine what degree of conviction should be associated with a claim once a piece of evidence has been received: prior degree of belief in the claim, how likely the evidence would be if the claim were true, and how likely it would be if the claim were false. Specifically, Bayes' theorem states that:

$$P(h | e) = \frac{P(h)P(e | h)}{P(h)P(e | h) + P(\neg h)P(e | \neg h)} \quad \text{Eq. 1}$$

where $P(h|e)$ represents one's posterior degree of belief in a hypothesis, h , in light of the evidence, e , which can be calculated from one's initial, prior degree of belief, $P(h)$, and how likely it is that the evidence one observed would have occurred if one's initial hypothesis was true, $P(e|h)$, as opposed to if it was false, $P(e|\neg h)$. The ratio of these latter two quantities, the likelihood ratio, provides a natural measure of the diagnosticity of the evidence, that is, its informativeness regarding the hypothesis or claim in question.

We will discuss the application of Bayes' theorem to individual arguments in more detail later, and we note here simply several general characteristics. First, what values any of these quantities take depends on what the statements in question are about, that is, the specific *content* of hypothesis and evidence. Second, the Bayesian framework, through its interpretation of probabilities as subjective degrees of belief, accords with the general intuition that argumentation contains an element of audience relativity (a property widely perceived to be characteristic of real-world arguments, see, e.g., Hahn & Oaksford, 2006a, 2006b; Perelman & Olbrechts-Tyteca, 1969; Toulmin, 1958); that is, the same argument need not (rationally) have the same impact on all recipients. Nevertheless, Bayesian probability imposes real constraints on the beliefs an agent can have, both by guaranteeing probabilistic consistency between the beliefs of a given agent, and because the beliefs of different agents are, in certain cases, guaranteed to converge

as these agents observe more and more evidence (see also, e.g., Hahn & Oaksford, 2006b).

There has been (and continues to be) debate about the proper scope of the procedural and epistemic frameworks just discussed. They target different aspects of argumentation, but both in theory and practice these aspects are intertwined. Dialectical and epistemic concerns are related. For example, silencing opponents by force is undoubtedly a violation of dialectical, procedural norms for "good" argumentation; but it seems dubious even for those interested not in process, but only in truth, because the suppression of arguments in discourse means that the potentially strongest argument might not be heard (see also Goldman, 1994; Hahn & Oaksford, 2006b). Likewise, pragma-dialectical theories have used discourse rules to evaluate fallacies of argumentation (e.g., van Eemeren & Grootendorst, 1992, 2004; Walton, 1995). However, the problem that remains is that discourse rules typically do not provide enough constraints on content. It is not hard to find examples of arguments with the same structure, and in the same argumentative context, that nevertheless differ fundamentally in how intuitively compelling they seem, and this has been at the heart of recent criticisms of the pragma-dialectical approach to the fallacies (e.g., Hahn & Oaksford, 2006a). Hence, normative theories of content and procedural theories can (as we will see) clash on occasion, but they ultimately pursue complementary goals (Goldman, 1994; Hahn & Oaksford, 2006b), and both have an important role to play in shaping "rational debate."

The Psychology of Argumentation

As detailed in the Introduction, psychological research has addressed argumentation from a number of different perspectives. In the following sections our main emphasis will be on basic research concerning argument quality; then in the remainder, we will provide brief introductions to research concerned with the development of argumentation skills, and educational attempts to foster argumentation, as well as argument in a number of specific practical contexts such as science and the courtroom.

Argument Quality

Foundational research on argument quality within psychology can itself be split naturally into research concerned with procedural aspects of argumentation, in particular pragma-dialectic

norms governing rational debate, and into research concerned with the epistemic quality of the argument, and hence, the actual main substance of its content.

PROCEDURAL ASPECTS OF ARGUMENTATION

Experimental research on procedural aspects of argumentation stems from a number of sources: cognitive psychologists, educational psychologists, communication researchers, argumentation theorists, and philosophers. This diversity of disciplines and theoretical backgrounds means that relevant psychological research is found beyond the confines of mainstream psychological outlets. In the following, we provide key examples.

A central topic for cognitive psychologists with an interest in pragma-dialectical aspects of argument has been the burden of proof (e.g., Bailenson, 2001; Bailenson & Rips, 1996; Rips, 1998; see Rips, Brem & Bailenson, 1999 for reviews). As noted earlier, the notion is derived from law, where burdens of proof are explicitly specified. In the context of psychological research, the notion has been operationalized by presenting participants with argumentative dialogues and asking them to indicate which proponent in the dialog “has the most to do in order to prove that he or she is right.” Experimental manipulations involve providing a proponent’s challenge (“What is your evidence for that statement?”) earlier or later in dialogue, and whether a dialogue starts with a neutral claim (“Hi, good to see you”) and then an assertion by the second speaker, or directly with the assertion of the second speaker. Such manipulations are found to have an effect; however, it is debatable to what extent these tasks involve a burden of proof in any technical sense (see Hahn & Oaksford, 2007b). It is clear that the evaluation of a series of arguments can be influenced by the order in which the component arguments are presented (see also McKenzie, Lee, & Chen, 2002). Crucially, order will affect the interpretation of material. For example, the order in which issues are put forward by a speaker are likely to be influenced by their perceived importance. This consideration allows corresponding inferences on the part of the listener, for example, about what the speaker is most concerned about. Likewise, changes in the order in which arguments and counterarguments appear can alter the perceived relevance of claims, and the degree to which they seem responsive to what has preceded them. Not seeking to refute a prior claim can be taken to signal tacit acceptance (see e.g., Clark, 1996), or at least

the lack of a cogent counterargument, in the same way that the provision of weak (counter-) evidence can be taken to suggest that no stronger evidence is available. These latter inferences are (nonfallacious) examples of argument from ignorance, an argument form we turn to in detail later (see Harris, Corner, & Hahn, 2009). At the same time, “diffuse” responses may be seen to affect the perceived competence of the person supplying them (Davis & Holtgraves, 1984) or their credibility (O’Keefe, 1999). Social psychologists have studied extensively the effects of so-called one-sided versus two-sided messages; that is, the effectiveness of ignoring versus acknowledging counterarguments. O’Keefe (1999) provided a meta-analysis of the respective persuasive effects of such one- versus two-sided messages (for a review of studies on this topic, see also Allen, 1991). Based on a systematic evaluation of over 50 studies, O’Keefe concluded that two-sided arguments that address counterarguments are typically more persuasive than one-sided arguments; least persuasive are arguments that explicitly acknowledge counterarguments without trying to refute them. From a pragma-dialectical perspective, this ordering ties in with the fundamental procedural obligation to defend one’s position when challenged (see earlier) in that the most persuasive arguments are those that at least attempt to discharge argumentative burdens (see also, O’Keefe, 2003); however, it is not clear that these are ultimately anything other than argument *content* effects (see also O’Keefe, 1999, for discussion of this point).

As part of an ongoing project to consider the extent to which the normative quality of arguments squares with their actual persuasive effect as established within social psychological research, O’Keefe has also conducted a meta-analysis of the persuasive effects of standpoint explicitness, that is, the extent to which the degree of articulation with which a communicator presents his or her overall conclusion affects message effectiveness (O’Keefe, 1997a). From a pragma-dialectical perspective, such explicitness can be linked to procedural obligations of the proponent of an argument to avoid “concealment, evasion, and artful dodging” and present information in such a way that allows critical scrutiny. In line with such obligations, social psychological research has found that better articulation seems to give rise to greater persuasive effect. O’Keefe’s (1997b) meta-analysis revealed corresponding effects for the degree of articulation in the actual argumentative support. In short, O’Keefe’s findings demonstrate

some correspondence between what might be considered normatively desirable and what, in persuasive terms, “actually works.”

Finally, procedural norms have been empirically investigated within the framework of “argumentational integrity” or fairness (e.g., Christmann, Mischo, & Groeben, 2000; Mischo, 2003; Schreier, Groeben, & Christmann, 1995). Here, studies have sought to examine people’s sensitivity to the violation of procedural rules, such as, for example, that proponents within a debate should have equal opportunity to contribute, or that contributions to debate must be sincere. Research has also examined the persuasive costs of such rule violations, and it has sought to develop educational training programs to increase awareness of violations (Christmann, Mischo, & Flender, 2000).

PISTEMIC ASPECTS OF ARGUMENT QUALITY

Logic and probability theory combine to provide powerful tools for the evaluation of arguments. Classical logic provides minimum standards by enforcing logical consistency and the avoidance of contradictions: A statement that is contradictory can never be true, and thus it constitutes neither a claim nor evidence worth consideration. Probability theory then constrains content beyond these minimal, logical requirements (see Chater & Oaksford, Chapter 2).

Both people’s logical reasoning and their ability to deal appropriately with probabilities have been the focus of vast amounts of psychological research (for overviews, see e.g., Hardman, 2009; Manktelow, 2011). Broadly construed, all of this research is relevant. Psychological work on logical reasoning, for example, deals with very specific and very restricted “arguments.” The logical reasoning literature has often been scathing about people’s logical abilities; however, there is reason to question the applicability of much of this research to everyday argumentation. For one, it is well documented that people’s degree of conformity to logic is much affected by the exact content of materials (Manktelow, 2011; Wason & Johnson-Laird, 1972; see Evans, Chapter 8). It has been argued recently that people’s abilities in this regard are much better when they are embedded in the kinds of argumentation contexts that logical reasoning supports in everyday life (Mercier & Sperber, 2011). Furthermore, many tasks that experimenters perceive to involve deduction may not necessarily do so from the perspective of the participant. In particular, it has been argued that people typically

adopt a probabilistic interpretation of conditionals (if-then statements), and once this is taken into account their reasoning is far from bad (e.g., Evans & Over, 2004; Oaksford & Chater, 1994, 2007, 2009). This issue is discussed extensively elsewhere in this volume (Chater & Oaksford, Chapter 2), but we will also consider several types of conditional. Analogous points apply to the literature on probability judgment. For one, people argue primarily about things they care about and with which they have some degree of familiarity. Moreover, although some evidence, and hence argument, involves overt numbers, probabilities, and statistics (and hence limitations identified in previous research may readily apply), most everyday argument does *not* cite numerical quantities. Consequently, for most of the many different argument forms that arise in everyday discourse, experimental research has only just begun.

Much of that research has been centered around putative cases of human error, that is, fallacies in human reasoning. Hence, we will focus primarily on these fallacies in the following sections, before concluding with research addressing argument quality in a number of applied contexts.

Fallacies of Argumentation: A Litmus Test for Evaluation of Argument Quality

The fallacies have long been a focal point for philosophical research on argument quality. There is debate about how to best define the notion of fallacy (see, e.g., van Eemeren & Grootendorst, 2004, for discussion). On an intuitive level, fallacies are simply arguments that might *seem* correct but aren’t, that is, arguments that might persuade but really should not. Contemporary lists include more than 20 different fallacies (see e.g., Woods, Irvine, & Walton, 2004). The fallacies of circular argument (Walton, 1985, 1991) and the argument from ignorance (Walton, 1992a) have already been mentioned; other well-known fallacies are the slippery slope argument (“if we legalize assisted suicide, it will be euthanasia next;” Walton, 1992b), the ad populum argument or appeal to popular opinion (“many people think this; it cannot be wrong;” Walton, 1980, 1999), the ad hominem argument, which seeks to undermine the proponent of an argument instead of addressing the argument itself (e.g., Walton, 1987, 1998b, 2000, and references therein), the ad verecundiam argument also known as the appeal to authority (e.g., Walton, 1997), equivocation (e.g., Engel, 1986; Kirwan, 1979; Walton, 1996b; Woods & Walton, 1979), and the

ad baculum argument, which appeals to threats or force (“if you don’t like that, you might find yourself out of a job;” Walton, 2000b; Walton & Macagno, 2007). These informal arguments are pervasive in everyday discourse (for real-word examples, see also, e.g., Tindale, 2007), and critical thinking textbooks seek to educate about them. The goal of philosophical research on the fallacies has been to provide a comprehensive treatment—ideally a formal treatment—of these fallacies that can explain exactly why they are “bad” arguments (see e.g., Hamblin, 1970). In other words, the fallacies are a litmus test for our theories of argument quality.

The fallacies have also been investigated experimentally from both a broadly procedural, typically pragma-dialectical perspective (e.g., van Eemeren, Garssen, & Meuffels, 2009; Neuman, 2003; Neuman, Glassner, & Weinstock, 2004; Neuman, Weinstock, & Glasner, 2006; Neuman & Weitzman, 2003; Ricco, 2003; Rips, 2002), and from an epistemic perspective (e.g., Hahn & Oaksford, 2007a; Oaksford & Hahn, 2004; Oaksford & Chater, 2007, 2010a,b). Although in many ways such experimental work has only just started, the general finding has been that people seem quite competent at identifying fallacious arguments.

A more detailed examination of both theory and experimental work on the fallacies also demonstrates why procedural approaches to argument quality are not sufficient. This is well illustrated with one of the more widely studied arguments, the argument from ignorance:

Ghosts exist, because nobody has proven that they don’t. (1)

Historically, one of the main stumbling blocks for theoretical treatments of the fallacies was that most of the fallacies seem to involve seeming “exceptions” in the form of examples that do not seem as bad (see e.g., Walton, 1995). The following examples too are arguments from ignorance, but unlike (1) seem acceptable:

This drug is safe because clinical trials have found no evidence of side effects. (2)

and

The book is in the library, because the catalog does not say that it is on loan. (3)

Clearly, the inferences in all three of these cases are uncertain or defeasible; that is, the conclusions do not follow necessarily from the evidence.

However, they seem sufficiently likely to be true, in light of the reason or evidence provided, that we readily base our actions on such arguments in everyday life. Examples such as (2), in particular, are widespread in socioscientific debate about the acceptability of new technologies (e.g., genetically modified foods, nanotechnology). Uncertainty is, of course, also present in positive inferences such as:

This drug causes side effects, because some participants in the clinical trial exhibited flu-like symptoms. (4)

Pragma-dialectical theories seek to explain why textbook examples of the argument from ignorance are poor arguments by considering them in a wider dialectical context, and attributing their insufficiency to the violation of discourse rules within that wider, overall argument. Specifically, arguments such as (1) are typically assumed to violate the burden of proof (see e.g., van Eemeren, Garssen & Meuffels, 2009; Walton, 1992a, 1995, 1996a). As we saw earlier, the burden of proof is assumed to demand that whoever makes a claim has to provide reasons for this claim when challenged. Stating that no one has disproved the existence of ghosts as a reason for believing in them constitutes an illegitimate attempt to shift that burden onto the other party, instead of providing an adequate reason oneself.

However, such an explanation seems forced for two reasons (see also Hahn & Oaksford, 2006a, 2007a, 2007b). First, example (1) seems intuitively a weaker argument than (2) or (3) even when all are presented in isolation as they are here, that is, without any wider argumentative context and any indication of parties engaged in debate. Here it is unclear to whom obligations such as burdens of proof could be ascribed. Second, violations of one’s burden of proof are a *consequence* of providing insufficient evidence, not a *cause*. The judgment that an argument seeks illegitimately to “shift the burden of proof” does not explain an argument’s weakness; rather, it presupposes it. Weak arguments fall short of burdens of proof; strong ones do not. Consequently, it still needs to (independently) be determined why, for example, (1) is poor, in ways that (2), (3), and (4) are not.

The identification of very abstract relations such as “claim,” “warrant,” or “backing” as found in Toulmin’s and similar systems also provides no guidance here. All four examples involve a claim and a reason given in support. Rather, an epistemic standard aimed at the specific content is required

here. Logic has nothing to say about these arguments; none of them are deductively valid.⁵

The probabilistic Bayesian framework, however, does make appropriate distinctions here (Hahn & Oaksford, 2006a; Hahn & Oaksford, 2007a; Oaksford & Hahn, 2004). The standard version of Bayes' theorem provided in Eq. 1 earlier applies directly to positive arguments such as (4). The claim that "this drug causes side effects" takes the place of hypothesis h ; the reason "because some participants in the clinical trial exhibited flu-like symptoms" takes the role of evidence e . This evidence is convincing to the extent that these symptoms are more likely if the claim is true, $P(e|h)$, than if it is not, $P(e|\neg h)$, for example, because it is winter and participants can catch a flu. As noted earlier, it is the specific content of the argument that fixes these probabilistic quantities, and argument strength will vary with the likelihood ratio (diagnosticity of the evidence). Observing, for example, that 95% of the participants in the trial displayed side effects within hours of taking the drug will provide greater support than observing a few who fell ill over a several-week period. Crucially, this approach allows one to capture content specific variation in the perceived strength of arguments of the same structure.

A negative argument such as (2) requires the corresponding negative form of Bayes' theorem:

$$P(\neg h | \neg e) = \frac{P(\neg e | \neg h)P(\neg h)}{P(\neg e | \neg h)(\neg h) + P(\neg e | h)P(h)}$$

Eq. 2

Again, such negative arguments can be weaker or stronger: Failing to observe side effects in 50 studies provides far stronger evidence for the claim that the drug lacks side effects (is safe) than does observing no side effects in just a single study, and these differences are readily captured.

However, the Bayesian framework also identifies important consequences of differences in structure. Formal analysis reveals that across a broad range of possible (and in everyday life plausible) numerical values for both how likely the evidence would be if the claim were true and if it were false, positive arguments are stronger than their corresponding negative counterparts based on the same set of values (Hahn & Oaksford, 2007a; Oaksford & Hahn, 2004). This observation provides an explanation for why arguments from ignorance are typically less

convincing than corresponding arguments from positive evidence. In other words, the framework captures both characteristics of particular argument types, and of particular instantiations of these types.

From a formal perspective, there is also an important difference between different types of argument from ignorance, exemplified on the one hand by the ghosts example in (1), and, on the other, by the drug safety example in (2). Whereas the just-discussed drug safety example simply involves an inference from a negative (lack of observed ill effects in clinical trials) to a negative (lack of side effects of drug), the ghosts example involves an inference from a double negation ("no-one has proven that they don't") to a positive (ghosts exist). This so-called epistemic closure case of the argument from ignorance (see, e.g., Walton, 1996a) requires a further distinction, because people could fail to prove that ghosts don't exist not only because they tried and actually found evidence of existence, but also because they did not try at all. Hence, Hahn and Oaksford's (2007a) Bayesian formalization of this type of argument from ignorance involves three possibilities: a report of positive evidence " e ," an explicit report of negative evidence " $\neg e$," and the third possibility, n (as in "nothing"), indicating that there is simply no evidence either way (i.e., neither an explicit reporting of e or $\neg e$). Such an approach is familiar from artificial intelligence where one might distinguish three possibilities in a database search regarding a proposition (h): the search can yield affirmative information (e), negative information ($\neg e$), or return with nothing (n). Epistemic closure has been invoked here to license inferences from search failure (i.e., a query resulting in nothing) to nonexistence, given that the database is assumed to be complete.

The Bayesian formalization of epistemic closure is analogous, except that closure can be a matter of degree, ranging from complete closure, through moderate closure, to no closure at all. The corresponding version of Bayes' theorem for the ghosts example is, therefore,

$$P(h | \neg'' \neg e'') = \frac{P(\neg'' \neg e'' | h)P(h)}{P(\neg'' \neg e'' | h)P(h) + P(\neg'' \neg e'' | \neg h)P(\neg h)}$$

Eq. 3

where $P(\neg'' \neg e'' | h) + P(\neg'' \neg e'' | \neg h) + P(n | h) = 1$.

How strong this argument is depends critically on the degree of closure. If the source is epistemically closed (i.e., the database complete), then the probability of a “nothing” response is zero, and everything reduces to the familiar binary case, where $\neg\neg e$ is the same as e (and one is effectively back to Eq. 1). As epistemic closure decreases, the probability of a null response increases, and the inference must become less strong (assuming that the quality of the explicit evidence we could obtain remains the same, that is, equally diagnostic; Hahn & Oaksford, 2007a, 2008). This fact explains why some cases of this argument, such as the library case in (3), are so much better than the ghosts example in (1). Electronic library catalogs are reasonably reliable, in that when they say a book is in the library, more often than not it actually is, because libraries try hard to keep their catalogs up to date. Likewise when catalogs say that a book is on loan, it typically is, and epistemic closure is high because loans are (in principle) always recorded.

Several experimental studies have examined the extent to which people’s judgments about arguments from ignorance are concordant with Bayesian prescriptions (Corner & Hahn, 2009; Hahn, Harris, & Corner, 2009; Harris, Corner, & Hahn, 2009; Hahn, Oaksford, & Bayindir, 2005; Hahn & Oaksford, 2007a; Oaksford & Hahn, 2004). Oaksford and Hahn (2004) provided participants with short dialogs between two characters, asking them to provide evaluative judgments of how convinced of a claim one of the characters should be in light of the evidence presented by the other. Two different scenarios were used, one involving drug safety and one involving TV violence. Participants saw both positive arguments (as in (4) above) and corresponding arguments from ignorance (such as (2) above). Also manipulated were the degree of prior belief the character in receipt of the evidence already had in the claim ($P(h)$), and the amount (and hence diagnosticity) of evidence received (e.g., one versus 50 clinical studies). Oaksford and Hahn found the expected main effects of prior belief, amount of evidence, and whether the argument was positive or an argument from ignorance. Participants saw the arguments from ignorance as acceptable but less acceptable than their positive counterparts, and that degree of acceptability was influenced by the prior beliefs of the characters in the dialog and the amount of evidence in the way predicted by a Bayesian account.

That participants distinguish clearly between what should be strong and weak versions of the argument from ignorance was confirmed in further

studies. For example, Hahn and Oaksford (2007a, Exp. 3) presented participants with the classic textbook example of the ghosts (1) and the library example (2), both embedded in short dialogs as in Oaksford and Hahn (2004). Participants’ ratings of how convinced a character in that dialog should be by the argument in question was significantly lower for the ghosts argument. Hahn et al. (2005) presented participants with positive arguments, as well as both types of argument from ignorance, and a negative argument involving explicit negative evidence. Furthermore, there were four different scenarios chosen to vary intuitively in the degree of epistemic closure. Participants’ evaluations were sensitive not only to the individual argument structures but also to the variations in closure, and—as Hahn and Oaksford (2007a) show—are well fit by the respective forms of Bayes’ theorem.

Finally, Harris et al. (2009) examined how silence or “nothing” itself can support rational inference that takes the form of arguments from ignorance. Imagine the recipient of a reference request in the context of a job application, who might be informed merely that the applicant is “punctual and polite” and, on the basis of that, conclude that the applicant is poorly qualified for the job—the phenomenon of being “damned by faint praise.” Here, “punctual and polite” constitutes a positive piece of evidence, which should (marginally) increase favorable evaluation of the candidate. What is doing the damage is what is not being said, namely, the absence of any discussion of job-relevant skills. Such an inference from nothing, n , to the conclusion that the applicant is not qualified ($\neg h$) is governed by this version of Bayes’ theorem:

$$P(\neg h | n) = \frac{P(\neg h)P(n | \neg h)}{P(\neg h)P(n | \neg h) + P(h)P(n | h)}$$

Eq. 4

and is licensed wherever $P(n|h) < P(n|\neg h)$, that is, wherever the probability of a “nothing” response is less if the hypothesis is true than if it is false. In this case, the nonoccurrence itself is informative because it is suggestive of the fact that the hypothesis is false. Hence, the effect should be observed where a motivated (or positively inclined) but nonlying source is presenting an argument. By contrast, there is no reason for this negative inference in the case of a maximally uninformed source, $P(n|h) \approx P(n|\neg h)$, who simply knows nothing on the topic.

In line with this, Harris et al.'s (2009) experiment involving a fictitious academic reference found that being told that "James is punctual and polite" led to decreases in judgments of whether James should be admitted to a university course in mathematics, but only when it came from his math tutor (who was presumably well informed about his mathematical abilities), but not from his personal tutor (who had provided general guidance only). Moreover, it was not the case that punctuality and politeness were perceived to be negative themselves, as when following details about mathematical ability this information raised evaluations of James' suitability.

A further case of inference from "nothing," finally, is the argument from ignorance we mentioned earlier: The failure to provide counterarguments or evidence can be taken to indicate either tacit agreement, or, at least, that the opponent has no counterevidence available. In this case, $P(n|h) > P(n|\neg h)$, and the lack of counterevidence suggests the claim is true. Moreover, the more salient a proponent has made a claim, thus indicating its perceived importance (for example, by introducing it early on), the more motivated the opponent should be to counter it, if possible. Hence, the more $P(n|h)$ should be assumed to exceed $P(n|\neg h)$, and thus the stronger the overall inference will be. It is in this way that changes to presentation order, or one- versus two-sided presentation, can directly affect the perceived content of the overall argument at hand.

CONDITIONALS

A probabilistic approach also deals naturally with arguments and fallacies that arise with the conditional (if...then) in natural language. In this area of the psychology of reasoning, an explicitly epistemic approach has emerged that is closely related to the research on argumentation we have so far reviewed (e.g., Oaksford & Chater, 2007, 2010a,b).

Conditional Fallacies

Many accounts of the fallacies include deductive fallacies such as those attaching to the conditional, if p (antecedent), then q (consequent). Two such fallacies are denying the antecedent and affirming the consequent. These fallacies are typically included among those that require an explanation in terms of the discourse rules involved in their use in argumentation (e.g., Godden & Walton, 2004). An example of denying of antecedent (DA) is:

If a bird is a swan, then it is white.
That bird was not a swan.
Therefore, that bird was not white. (5)

Godden and Walton's approach is to point out that while clearly a truth functional fallacy—as there are nonwhite swans—deploying this line of reasoning in argument against someone who is using this conditional to argue that a particular bird was white may be a sound strategy. Of course, if whether "that bird was white" is the topic of an argument, then it is clear that the parties in the argument disagree, and hence there is some uncertainty about the bird's color. The party deploying the conditional originally ("Pro") argues that the bird is white via a modus ponens (MP) argument:

(Pro) If a bird is a swan, then it is white. That bird was a swan; therefore, it was white.

To refute Pro's argument, the respondent ("Con") must deny one of the premises. Con chooses to "deny the antecedent," that is, to deny that the bird was a swan, from which it "follows" that the bird was not white:

(Con) But that bird was not a swan; therefore, it was not white.

However, as Godden and Walton observe, in this example, the falsity of the consequent—the bird was not white—does not follow from this use of denying the antecedent in the way that it would if it were logically valid, that is, it is not truth preserving. Rather, its deployment here undermines another property of Pro's intended conclusion, what "might be called its *admissibility*, or that it follows or that it is established or that it may be concluded, or perhaps even that it should be believed" (Godden & Walton, 2004, p. 232). A subjective probabilistic approach to argument strength for the conditional can cash out this last intuition with respect to whether Pro's conclusion is believable (Oaksford & Hahn, 2007).

According to Hahn and Oaksford's (2006a, 2007a) account of argument strength, people are concerned with the subjective degree of belief they should have in the conclusion given their degrees of belief in the premises. Oaksford and Chater (2007; Oaksford, Chater, & Larkin, 2000) proposed just such a model for the conditional. We will not rehearse the theory in any detail. It depends on people possessing prior degrees of belief in the conditional, given by the conditional probability, $P_o(q|p)$, and in the marginals, $P_o(p)$, $P_o(q)$ (the "0"

subscript denotes prior probabilities; a “1” subscript denotes posterior probabilities). These probabilities define a prior probability distribution from which, for example, the probability $P_0(\neg q|\neg p)$ can also be derived. People are assumed to derive their degrees of belief in the conclusions of conditional inferences by Bayesian conditionalization. So, for modus ponens, when Pro asserts that “that bird was a swan,” for Bayesian conditionalization to apply, she is asking her audience (here Con) to assume that $P_1(p) = 1$, from which it follows that one’s degree of belief in the conclusion, $P_1(q)$, should be $P_0(q|p)$.

It is straightforward to derive contexts in which denying the antecedent is stronger than modus ponens. So if Con’s beliefs lined up pretty closely to such a context, his counterargument by denying the antecedent could be stronger than Pro’s initial argument by modus ponens. Suppose that the swans they are talking about are in a bird sanctuary containing equal numbers of white and black swans, and that most of the birds in the sanctuary are neither white nor swans. The following distribution captures these facts: $P_0(q|p) = .5$, $P_0(p) = .1$, $P_0(q) = .1$. On this distribution, a bird is nine times more likely to be white given it is a swan than that it is not a swan. However, the probability that the bird is white given it is a swan is only .5, i.e., $P_1(q) = .5$. That is, on this distribution, Pro’s argument while logically valid, only yields a .5 degree of belief in the conclusion. Con’s argument can be characterized as noting that priors matter and that it is highly unlikely that the bird was a swan; that is, Pro’s assertion that $P_1(p) = 1$ is unlikely to be true. For Bayesian conditionalization to apply, this has to be incorporated into an argument as the assumption that $P_1(p) = 0$, and so $P_1(\neg p) = 1$. The posterior probability is $P_1(\neg q) = P_0(\neg q|\neg p) = .94$. That is, on this distribution, Con’s DA argument is stronger than Pro’s MP argument, in the sense of yielding a higher degree of belief in its conclusion.

One may also avoid the fiction of Con (or Pro) asking his or her interlocutor to assume that the categorical premise is certain, $P_1(\neg p) = 1$. By using Jeffrey conditionalization—a generalization of Bayesian conditionalization to when the categorical premise is uncertain (see, e.g., Jeffrey, 2004)—one could just argue that the probability that the bird was a swan equals one’s subjective degree of belief, that is, .1, then $P_1(\neg q) = P_0(\neg q|\neg p) P_1(\neg p) + P_0(\neg q|p) P_1(p) = .94 \times .9 + .5 \times .1 = .896$. This reformulation does not change things by very much, that is, DA is still the stronger argument, and we should believe

its conclusion more than the conclusion of the MP argument. So by developing an account of argument strength using subjective probability, we can generate a measure of how much the conclusion of the DA argument “*should be believed*.” It remains for Con to persuade Pro that the distributions on which the former bases his argument map on to the way the world actually is.

We then confront an issue we have raised before (Hahn & Oaksford, 2007a): What comes first, the assessment of the strength of the respective arguments given what Con believes, or the deployment of the DA argumentative strategy? It seems clear that deploying DA in this context is appropriate because Con believes it to be the stronger argument. The burden of proof of course then returns to Con to establish that the context is as Con believes it to be.

However, there are features of this context that suggest that one might rarely be justified in deploying DA in argument. First, ignoring the priors, Pro’s MP argument is more forceful. It should lead to greater change in people’s degree of belief in the conclusion. This is because the bird is 9 times more likely to be white given it is a swan but only 1.88 times more likely not to be white given it is not a swan. Oaksford and Hahn (2007) proposed the likelihood ratio as a possible prior independent measure of argument *force*. Ignoring priors is a well-documented phenomenon in the psychology of judgment and decision making (Bar-Hillel, 1980; Kahneman & Tversky, 1973). Hence, it is understandable why despite the low prior, Pro (even if he believed the prior to be low) might view MP as a “good” argument to put forward. Con’s DA counterargument is in a sense a reminder that the priors do matter, and that in this case one should be more convinced that the bird is not white.

Second, however, one might question whether there are many circumstances in which the DA argument is justified because it is the stronger argument. One of the assertability conditions on the conditional, at least according to Adams (1998), is that $P_0(q|p)$ is high, certainly greater than .5. When this is the case, MP will usually be a stronger argument than DA, that is, $P_0(q|p) > P_0(\neg q|\neg p)$. So, by introducing the conditional, Pro may be implicitly asserting that MP is stronger than DA. This situation would seem to warrant a different argumentative strategy on Con’s part, that is, denying the conditional premise rather than the antecedent. Moreover, in the psychology of belief revision it has been found that in response to a contradiction of

the conclusion of an MP inference, people choose to revise their belief in the conditional rather than their belief in the categorical premise (Elio & Pelletier, 1997). In arguing that Pro—or Pro and Con's audience—should believe the opposite of the conclusion of the MP inference, Con is in a similar position and hence may normally choose to deny the conditional rather than the antecedent.

Deontic and Utility Conditionals

In the psychology of reasoning, it has been observed that conditional sentences frequently describe situations that have associated utilities, which may figure in various argumentative strategies and fallacies (Manktelow & Over, 1987). This observation was first made in the context of deontic reasoning (Cheng & Holyoak, 1985; Cosmides, 1989), that is, about what you should and should not do, using conditionals like,

If you are drinking beer, you must be over
18 years of age. (6)

An enforcer of such a deontic regulation will place a high utility on detecting cheaters, that is, people who are drinking beer but who are not over 18 years of age. Oaksford and Chater (1994) showed how people's behavior on the deontic version of Wason's selection task could be captured by assuming that people are attempting to maximize expected utility in their selection behavior. More recently, Perham and Oaksford (2005) showed how this calculation could be modified by emotions to explain effects when explicitly threatening words were used in the antecedent of a deontic conditional. Bonnefon (2009) has developed a classification scheme for utility conditionals where the positive or negative utility is associated with either the antecedent or consequent of a conditional, and there have been several papers investigating such effects (Evans, Neilens, Handley, & Over, 2008; Thompson, Evans, & Handley, 2005).

Evans et al. (2008) investigated a variety of conditionals expressing conditional tips, warnings, threats, and promises. For example, "If you go camping this weekend, then it will rain" is a clear warning not to go camping. The higher $P(q|p)$ and the more negative the utility associated with the consequent, $U(q)$, that is, rain, the more persuasive such a conditional warning is to the conclusion that action p should not be taken, $\neg p$, that is, you should not go camping. For warnings, Evans et al (2008) argued that the decision about whether to perform

the action p is based on the *prima facie* utility of the action itself, $U(p)$, less the expected disutility of the action to which p could lead, $P(q|p)U(q)$, that is,

$$U(p) - P(q|p)U(q) \quad \text{Eq. 5}$$

Hahn and Oaksford (2007a) provided a very similar analysis of the slippery slope argument (SSA), which has often been regarded as an argumentative fallacy.

Slippery Slope Arguments

SSAs are typically expressed in conditional form (see Corner, Hahn, & Oaksford, 2011):

If we allow gay marriage, then in the future
people will want to marry their pets. (7)

If voluntary euthanasia is legalized, then in
the future there will be more cases of "medical
murder." (8)

If we accept voluntary ID cards in the UK, we
will end up with compulsory ID cards in the
future. (9)

These examples have all actually been put forward in the media by groups arguing that the antecedent actions of (7) to (9) should not be taken. As these examples show, SSAs can vary greatly in strength. Like conditional warnings, the conclusion people are invited to draw is $\neg p$, for example, one should not allow gay marriage. (7) is weak because of the very low value of $P(q|p)$, whatever we may think of the merits of interspecies marriage. (8) is stronger because this probability is higher but also because "medical murder" is clearly so undesirable, that is, $U(q)$ is strongly negative. (9) is even stronger because $P(q|p)$ seems very close to 1 and the consequent is highly undesirable (for some).

What differs between SSAs and warnings is that, whereas for warnings $P(q|p)$ is assessed just by reference to prior world knowledge, for SSAs there seems to be an implied mechanism that leads to the consequent action from the antecedent action. This mechanism suggests that an act of categorizing an item a (gay couples) under a category F (can marry), that is, F_a , will lead to other items b (interspecies "couples") also falling under the same category, F_b . Hahn and Oaksford (2007a) proposed that such a "category boundary re-appraisal" mechanism may explain why people find slippery slope arguments so compelling.

Specifically, current theories of conceptual structure typically agree that encountering instances of a

category at the category boundary should extend that boundary for subsequent classifications, and there is a wealth of empirical evidence to support this (see Rips et al., Chapter 11). In particular there are numerous experimental demonstrations of so-called exemplar effects, that is, effects of exposure to particular instances and their consequences for subsequent classification behavior (e.g., Lamberts, 1995; Nosofsky, 1986, 1988a, 1988b). For example, observing that a dog that weighs 10 kg is considered underweight invites the conclusion that a dog that weighs 10.5 kg is also underweight. With only the information that a 5 kg dog is underweight, and a 15 kg dog is overweight, however, one might not be so compelled to draw this conclusion. This is because of the similarity between 10 kg and 10.5 kg and the comparative dissimilarity with either 5 kg or 15 kg. Similarly, one may argue that (7) is a poor argument and so $\text{Pr}(q|p)$ is low because of the dissimilarity between same-sex human relations and interspecies relations, and that hence it is clear that there is no likelihood of slippage of the category “can marry” from one case to the other.

Corner, Hahn, and Oaksford (2011) have shown that people’s confidence in judging that various acts fall under a certain category is directly related to their willingness to endorse a corresponding SSA, and that this relationship is moderated by the similarity between the acts. For example, participants who are told that *assault in possession of a knife* has been categorized as having a tariff of *less than 20 years imprisonment* may confidently decide that *assault in possession of a gun* would also be given the same tariff. These same participants also endorse the slippery slope argument to the conclusion that *assault in possession of a knife* should be given a tariff of greater than 20 years because giving it less than 20 years may lead to *assault in possession of a gun* also being given the lower tariff. In this condition in Corner et al.’s (2011) Experiment 3, decision confidence in the categorization judgment and SSA endorsement were substantially correlated, $r(25) = .47$. When *assault without a weapon* is substituted for *assault in possession of a knife*, that is, an offense less close to *assault in possession of a gun* in similarity space, decision confidence and SSA endorsement rates become less correlated. In this dissimilar condition in Corner et al.’s (2011) Experiment 3, the correlation between decision confidence in the categorization judgment and SSA endorsement fell dramatically, $r(24) = .04$. The moderating effect of similarity was confirmed in a moderated regression analysis (Aguinis, 2004).

In sum, current work on conditionals and associated arguments shows that a Bayesian account of argument strength with associated utilities (SSAs) and without them (Denying the antecedent) provides a rational understanding of how these apparent fallacies of reasoning and argumentation function. The assessment of the conditional probability, $P(q|p)$, is central for assessing argument strength in both cases. For conditional inferences and fallacies, this is a reflection of world knowledge and can be assessed using the generic Ramsey Test for conditionals (Edgington, 1995; Ramsey, 1931). The Ramsey Test runs as follows: Add the antecedent, p , to your stock beliefs, make adjustments to accommodate this belief, and read off your degree of belief in the consequent, q and this then is $P(q|p)$. However, the Ramsey test is a philosophical prescription crying out for a psychological, algorithmic level explanation (Oaksford & Chater, 2007, 2010b,c). Constraint satisfaction processes in neural networks provide one possible implementation (Oaksford & Chater, 2010b). We have argued that for many SSAs, $P(q|p)$ is determined via category boundary reappraisal processes, which may represent a further algorithmic instantiation of the Ramsey Test (Corner, Hahn, & Oaksford, 2011; Hahn & Oaksford, 2007a).

CIRCULAR ARGUMENTS

The best-known fallacy within the catalog is circularity or “begging the question.” It is also the fallacy that has attracted the most attention and that, arguably, generates the most confusion. Different types of circular argument have been distinguished. One, often termed *equivalency circularity* (see e.g., Walton, 2005), takes the form “ A , therefore A .” This argument type has been theoretically puzzling to philosophers because it is logically valid, and thus in some ways “good;” but it is unlikely to occur very often in practice. A second type is the so-called *dependency circularity*, of which one example is so-called self-dependent justification as in “God exists because the Bible says so and the Bible is the word of God.” Here the evidence presupposes, and in that sense depends on, the very conclusion it is seeking to support; specifically, the Bible cannot be the word of God if God does not exist, which of course is the claim at stake. Though this example will strike most as a rather weak argument, it has been pointed out that many scientific inferences are also self-dependent in that they involve theory-laden observations, and that this self-dependence does *not* necessarily rule them out as legitimate arguments

(for actual scientific examples see Brown, 1993, 1994). This is revealed by a probabilistic analysis (e.g., Hahn & Oaksford, 2007a; Shogenji, 2000), which makes clear that self-dependent arguments as found frequently in science can be acceptable, because the conclusion can be held tentatively, and evidence diagnostic of that conclusion will, when actually observed, increase the posterior of degree belief we have in that conclusion (even though the conclusion itself is involved in the interpretation of the evidence).

There have been a number of psychological experiments examining people's awareness of circularity and their ability to distinguish stronger from weaker circular arguments (e.g., Hahn & Oaksford, 2007a; Rips, 2002). While such studies have generally found that people perform competently, there is clearly a limit here. Psychological researchers themselves frequently accuse others of circularity, and closer inspection suggests that this charge is often overused (Hahn, 2011). Moreover, the distinctions between different types of circular arguments are not always appreciated.

In fact, circular arguments provide a key demonstration of how useful and important the formal tools of probability theory are. Bayesian prescriptions often seem intuitively obvious, and there is a view that the theory of probabilities is "at bottom nothing but common sense reduced to calculus" (Laplace, 1814/1951). However, there are cases where our everyday intuitions break down. Circularity is a case in point. To provide one more example, we might find it intuitive that a sequence or chain of supporting evidence cannot be circular. This is to say that we might find it intuitive that a chain of evidence $E_0, E_1 \dots E_n$ where E_1 supports E_0 , E_2 supports E_1 , and so on (in the familiar sense of support $P(E_0|E_1) > P(E_0|\neg E_1)$) cannot ultimately loop back on itself and have E_n supported only by E_0 , that is, the very piece of evidence support was being sought for. However, Atkinson and Peijnenburg (2010) show mathematically that this is entirely possible, both for finite and infinite loops, and that the presence of such loops does not preclude the possibility that the probability of $P(E_0)$ be well defined. To illustrate with one of their examples:

A: Peter read parts of the philosopher Kant's magnum opus the "Critique of Pure Reason."

B: Peter is himself a philosopher.

C: Peter knows that Kant defended the synthetic a priori.

Here it is both possible and plausible that A is supported (i.e., made more probable) by B, which in turn is supported by C which itself is supported by A. Although one might intuitively suspect that the circularity here means that the (unconditional) probability assigned to A must always remain "up in the air," this is mathematically not so, and given an appropriate assignment of conditional probabilities ($P(A|B)$, $P(B|C)$, $P(C|A)$), a definite value for $P(A)$ ensues.

SUMMARY

The fallacies discussed so far are arguably the most prominent in the philosophical literature, but as noted earlier, they by no means exhaust the catalog of fallacies. The theoretical issues that the remaining fallacies raise parallel those discussed so far (see Hahn & Oaksford, 2006a), but the majority have not yet been the subject of psychological research. Moreover, argumentation theorists have identified many further, nonfallacious argumentation schemes (see Walton, Reed, & Macagno, 2008), which have not been examined by psychologists. Reasoning research within psychology has gradually started to move beyond the very narrow set of (deductive) paradigms that have dominated it for the last decades, and this wealth of different types of argument should provide both fertile and practically important ground.

Argumentation Applied

The practical importance of argumentation to our everyday lives is reflected in the fact that there has also been a wealth of psychological research on argumentation in applied contexts, and this body of research, if anything, exceeds that of the fundamental research we have described so far. Research into the development of children's argument skills is frequently closely tied to the educational goal of fostering and improving such skills (e.g., Anderson, Chinn, Chang, Waggoner, & Yi, 1997; Brem & Rips, 2000; Genishi & DiPaolo, 1982; Glassner, Weinstock, & Neuman, 2005; Kuhn, 1989, 1991, 2001; Means & Voss, 1996).

A considerable amount of this research has been concerned specifically with science (Kuhn, 1993). Much of it has been focused on tracking the development and quality of scientific reasoning (i.e., the use of hypotheses and evidence) in children (see, e.g., Klaczynski, 2000; Kuhn, Cheney & Weinstock, 2000; Kuhn & Udell, 2003; also Dunbar & Klahr, Chapter 35) and has been linked to educational

policy and programs designed to address deficits in scientific literacy (for reviews, see, e.g., Norris & Phillips, 1999; Sadler, 2004; see also Koedinger & Roll, Chapter 40).

A popular strategy in analyzing student use and recognition of different types of scientific argument has been to apply Toulmin's (1958) model. The ability to use (and recognize in other people) different aspects of argumentation such as data or warrants is used as an indicator of comprehension and rhetorical competence (e.g., Driver, Newton, & Osborne, 2000; Erduran, Simon, & Osborne, 2004; Jiminez-Aleixandre, 2002; Jiminez-Aleixandre, Rodriguez, & Duschl, 2000; Kortland, 1996; von Aufschaiter, Erduran, Osborne, & Simon, 2008). However, the limitations of this scheme discussed earlier have made themselves felt. In the words of Driver et al. (2000), "Toulmin's analysis...is limited as, although it can be used to assess the structure of arguments, it does not lead to judgments about their correctness..." (p. 294). Toulmin's model is a powerful tool for specifying the component parts of arguments and classifying them accordingly. It deals only, however, with the structure of different arguments, not their content.

Some researchers have developed their own, typically qualitative, evaluation schemes (e.g., Korpan, Bisanz, Bisanz, & Henderson, 1997; Kuhn, Shaw, & Felton, 1997; Norris, Phillips, & Korpan, 2003; Patronis, Potari, & Spiliotopoulou, 1999; Ratcliffe, 1999; Takao & Kelly, 2003), but these do not necessarily extend in applicability even to other studies.

In short, norms for the evaluation of argument content have been missed in this area. In fact, the Bayesian framework seems particularly suited to this context, as it has been so influential in the philosophy of science where it has been used to explain the ways in which scientists construct, test, and eliminate hypotheses, design experiments, and statistically analyze data (e.g., Bovens & Hartmann, 2003; Earman, 1992; Howson & Urbach, 1993). Corner and Hahn (2009) also demonstrate the utility of Bayesian probability as a heuristic framework for psychological research in this area. Specifically, they sought to examine whether there were systematic differences in the way participants evaluated arguments with scientific content (e.g., about genetically modified foods or climate change) and arguments about mundane, everyday topics (e.g., the availability of tickets for a concert). Although these arguments differ radically in content, they can be compared to each other, because both can

be compared to Bayesian prescriptions. Are people's evaluations more normative in one of these areas, and if yes, where do the systematic deviations and discrepancies reside? Once again, probability theory provides a tool for fine-grained analysis.

Finally, there has been considerable interest in argument and evidence within the courtroom (see Spellman & Schauer, Chapter 36). This is part of extensive psychological research into jury decision making involving mock juries. This research has often focused on the role of narratives or "stories" in legal argument. A story, in this sense, is a series of interrelated episodes that involves initiating events, goals, actions, consequences, and accompanying states (typically) in the order in which they occurred (e.g., Pennington & Hastie, 1981). Experimental research with mock juries has sought support for the claim that evidence is most naturally organized within such stories, and when presented in this way, is more memorable and has greater persuasive effects (see e.g., Pennington & Hastie, 1981, 1986, 1988, 1992; Voss & van Dyke, 2001).

A narrative structure is, however, not the only way in which legal arguments are made (see e.g., Schum, 1993), and, in terms of persuasive success not necessarily always the most effective (see Spiecker & Worthington, 2003). Other factors that have been examined are, for example, the impact of the comprehensiveness of opening statements by prosecution and defense and the extent to which these statements foreshadow subsequent arguments (Pyszczynski & Wrightsman, 1981; see also Pyszczynski, Greenberg, Mack, & Wrightsman, 1981). Many have examined issues surrounding possible bias through prior beliefs (e.g., Giner-Sorolla, Chaiken, & Lutz, 2002; on biased assimilation more generally (see Lord & Taylor, 2009; also see Molden & Higgins, Chapter 20); and there has been research into how argument evaluation and overall decisions interact (Carlson & Russo, 2001; Holyoak & Simon, 1999; Simon, Pham, Le, & Holyoak, 2001). Finally, there have been many studies concerned with different aspects of testimony (see e.g., Eaton, Ball, & O'Callaghan, 2001; Ivkovic & Hans, 2003; McQuiston-Surrett & Saks, 2009; Skolnick & Shaw, 2001; Weinstock & Flaton, 2004).

Normative concerns are integral to the legal process and the study of argument evaluation within law. However, law has, historically, tended to mistrust Bayesian probability and its appropriateness for capturing uncertainty in law (see, e.g., Tillers & Green, 1988); and psychological research

evaluating testimony relative to Bayesian prescriptions has been focused on quantitative, statistical evidence (in particular, probability estimates associated with biological profile evidence such as blood typing, hair fibers, and DNA; for a review see, e.g., Kaye & Koehler, 1991). Such studies reveal difficulties for many (mock) jurors in dealing with explicit numerical information. Beyond this, the dominant experimental finding here has been that participants underweight the evidence relative to Bayes' theorem. This finding is consistent with a long line of studies within social psychology examining quantitative aspects of belief revision (e.g., Edwards, 1968; Fischhoff & Beyth-Marom, 1983; but see also Corner, Harris, & Hahn, 2010; Erev, Wallsten, & Budescu, 1994). One contributing factor to this apparent deficit could be that participants treat the expert sources providing the evidence as less reliable than the experimenters assume. Schklar and Diamond (1999) specifically examined this possibility and found that basing calculations of "normative" posteriors on participants' own assessments significantly decreased the gap between normative and descriptive, although it did not remove it entirely.

The issue of testimony, of course, extends beyond just the courtroom. It can be argued that the majority of our knowledge depends, in one form or other, on the testimony of others (e.g., Spellman & Tenney, 2010). Hence, it is fundamental to argument evaluation that not just the content of an argument but also its source be taken into account. In general, the same evidence presented by a less reliable source *should* lead to a smaller change in degree of belief; moreover, characteristics of the argument or evidence itself and of the reporting source should interact multiplicatively from a normative, Bayesian perspective (see, e.g., Schum, 1981). Initial studies involving simple day-to-day arguments find support for the position that participants have some basic sensitivity to this general relationship (Hahn, Harris, & Corner, 2009; see that paper also for discussion of social psychological research on this issue).

That the utility of the Bayesian framework extends beyond overtly statistical contexts is also illustrated by recent work on the notion of coherence, that is, the extent to which multiple testimonies "fit together." Coherence is a central notion within the story model mentioned earlier, and recent work within Bayesian epistemology by the philosophers Stephan Hartmann and Luc Bovens has sought to understand how, when, and why informational coherence impacts belief change (Bovens & Hartmann, 2003).

A recent experimental investigation found that participants' judgments (regarding witness reports to the police on the putative location of a body) corresponded closely to Bayesian predictions, and clearly went beyond a simple averaging strategy (Harris & Hahn, 2009).

Finally, as was argued in the context of research on science education and communication earlier, the Bayesian framework can provide an invaluable methodological tool. Many of the questions that have been asked about testimony, such as whether it is more or less convincing than physical evidence (e.g., Skolnick & Shaw, 2001), or whether expert witnesses are more convincing when they provide qualitative or quantitative evidence (e.g., McQuiston-Surrett & Saks, 2009; for research on qualitative versus quantitative arguments more generally see also the reviews of Hornikx, 2005, 2007, and the studies of Hoeken & Hustinx, 2009) can only really be addressed if the specific *content* of the evidence is controlled for. It is in the nature of these particular research questions that it is typically not possible to exactly equate the content, and vary only the dimension of interest, thus leaving results inherently confounded. However, the Bayesian framework can contribute to such research because it allows such content-specific variation to be factored out (Corner & Hahn, 2009). Priors, posteriors, and likelihoods provide a common currency within which different types of arguments can be contrasted in search of presentation or domain-specific systematic differences.

Conclusions and Future Directions

Rather trivially, reviews of research literatures often end with calls for more research. However, in the case of argumentation, empirical research has, in many ways, genuinely only just commenced. The range of material discussed in this chapter has, to our knowledge, not even previously been brought together, and there are undoubtedly aspects where this synthesis remains incomplete. Procedural and epistemic norms for argument evaluation can now be brought together in order to develop a unified comprehensive treatment that was not previously possible. As outlined in the Introduction, argumentation is an inherently interdisciplinary field with enormous theoretical and practical interest. Psychology has a central role to play in this. Decades of social psychological research have found that argument quality seems the most influential factor in persuasion, but persuasion research has lacked the theoretical tools to address the question

of what makes an argument “good.” This issue has been identified time and again as the most serious gap in persuasion research (see, e.g., Fishbein & Ajzen, 1981, 2009; Johnson, Maio, & Smith-McLallen, 2005; O’Keefe, 1995). As detailed in this chapter, norms for argument quality exist; it remains to be explored in detail how sensitive people are to these.

Acknowledgments

Many thanks go to Adam Corner, Adam Harris, and Nick Chater for their role in helping shape our views on argument. Many thanks also to Jos Hornikx, Hans Hoeken, Lance Rips, and Keith Holyoak for helpful comments on earlier versions of the chapter.

Notes

1. It should be noted that Toulmin’s system is not the only, or even oldest, system like this. One of the most influential systems for representing dependencies between arguments is the Wigmore chart (for an introduction to this and other systems for displaying argument structure, see, e.g., Schum, 1994).

2. It is worth noting that there have always been connections between legal argument and other types of argumentation, in particular within the rhetorical tradition, for example, in Whately (1828), but Toulmin’s work gave considerable impetus to the view that legal argument constitutes a role model for rational debate more generally.

3. It is not enough for parties in a debate to simply agree upon a procedure or criteria for evaluation in order for the debate to qualify as rational (but see van Eemeren & Grootendorst, 2004, p. 151): basing argument evaluation on the reading of entrails, position of the stars, or drawing of lots, for example, would seem incompatible with rational debate even if both parties approved of these methods.

4. In its emphasis on probabilities, the Bayesian approach to argumentation has close links to early work on attitude change involving the subjective probability model, or “probabilogical model,” of cognitive consistency (McGuire 1960a, 1960b, 1960c; Wyer, 1970; Wyer & Goldberg, 1970), which has also had some application in argumentation studies (e.g., Allen, Burrell, & Egan, 2000; Hamble, 1977). The probabilogical model draws on the law of total probability to relate experimentally induced changes in one or more propositions to attendant changes in another (typically logically) related proposition (see also, Eagly & Chaiken, 1993, Ch. 5, for an introduction).

5. Nor are any of the many nonclassical logics described in Prakken and Vreeswijk (2002) helpful here. Though often aimed specifically at dealing with uncertainty, the support and consequence relationships they specify are, like Toulmin’s system, too undifferentiated to distinguish the four cases. In addition, desirable core properties of classical logic are typically lost.

References

- Adams, E. W. (1998). *A primer of probability logic*. Stanford, CA: CSLI Publications.
- Aguinis, H. (2004). *Regression analysis for categorical moderators*. New York: Guilford Press.
- Allen, M. (1991). Meta-analysis comparing the persuasiveness of one-sided and two-sided messages. *Western Journal of Speech Communication*, 55, 390–404.
- Allen, M., Burrell, N., & Egan, T. (2000). Effects with multiple causes: evaluating arguments using the subjective probability model. *Argumentation and Advocacy*, 37, 109–116.
- Anderson, R. C., Chinn, C., Chang, J., Waggoner, M., & Yi, H. (1997). On the logical integrity of children’s arguments. *Cognition and Instruction*, 15, 135–167.
- Aristotle. (2004). *On sophistical refutations* (W. A. Pickard-Cambridge, Trans.). Whitefish, MT: Kessinger Publishing Co.
- Atkinson, D., & Peijnenburg, J. (2010). Justification by infinite loops. *Notre Dame Journal of Formal Logic*, 51, 407–416.
- von Aufschaiter, C., Erduran, S., Osborne, J., & Simon, S. (2008). Arguing to learn and learning to argue: Case studies of how students’ argumentation relates to their scientific knowledge. *Journal of Research in Science Teaching*, 45, 101–131.
- Bailenson, J. (2001). Contrast ratio: Shifting burden of proof in informal arguments. *Discourse Processes*, 32, 29–41.
- Bailenson, J. N., & Rips, L. J. (1996). Informal reasoning and burden of proof. *Applied Cognitive Psychology*, 10, S3–S16.
- Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, 44, 211–233.
- Bench-Capon, T. M., & Dunne, P. E. (2007). Argumentation in artificial intelligence. *Artificial Intelligence*, 171, 619–641.
- Boger, G. (2005). Subordinating truth – is acceptability acceptable? *Argumentation*, 19, 187–238.
- Bonnefon, J. F. (2009). A theory of utility conditionals: Paralogical reasoning from decision theoretic leakage. *Psychological Review*, 116, 888–907.
- Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford, England: Oxford University Press.
- Brem, S., & Rips, L. (2000). Explanation and evidence in informal argument. *Cognitive Science*, 24, 573–604.
- Brown, H. (1993). A theory-laden observation can test a theory. *British Journal for the Philosophy of Science*, 44, 555–559.
- Brown, H. (1994). Circular justifications. *PSA*, 1, 406–414.
- Carlson, K. A., & Russo, J. E. (2001). Biased interpretation of evidence by mock jurors. *Journal of Experimental Psychology: Applied*, 7, 91–103.
- Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, 17, 391–416.
- Christmann, U., Mischo, C., & Flender, J. (2000). Argumentational integrity: A training program for dealing with unfair argumentative contributions. *Argumentation*, 14, 339–360.
- Christmann, U., Mischo, C., & Groeben, N. (2000). Components of the evaluation of integrity violations in argumentative discussions: Relevant factors and their relationships. *Journal of Language and Social Psychology*, 19, 315–341.
- Clark, H. H. (1996). *Using language*. New York: Cambridge University Press.
- Corner, A. J., & Hahn, U. (2009). Evaluating science arguments: Evidence, uncertainty & argument strength. *Journal of Experimental Psychology: Applied*, 15, 199–212.
- Corner, A., Hahn, U., & Oaksford, M. (2011). The psychological mechanism of the slippery slope argument. *Journal of Memory and Language*, 64, 133–152.
- Corner, A. J., Harris, A. J. L., & Hahn, U. (2010). Conservatism in belief revision and participant skepticism. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1625–1630). Austin, TX: Cognitive Science Society.
- Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, 31, 187–276.

- Davis, D., & Holtgraves, T. (1984). Perceptions of unresponsive others: Attributions, attraction, understandability and memory of their utterances. *Journal of Experimental Social Psychology*, 20, 383–408.
- Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education*, 84, 287–312.
- Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77, 321–357.
- Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Belmont, CA: Thompson/Wadsworth.
- Earman, J. (1992). *Bayes or bust?* Cambridge, MA: MIT Press.
- Eaton, T. E., Ball, P., & O'Callaghan, M. G. (2001). Child-witness and defendant credibility: Child evidence presentation mode and judicial instructions. *Journal of Applied Social Psychology*, 31, 1845–1858.
- Edgington, D. (1995). On conditionals. *Mind*, 104, 235–329.
- Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17–52). New York: Wiley.
- Elio, R., & Pelletier, F. J. (1997). Belief change as propositional update. *Cognitive Science*, 21, 419–460.
- Engel, S. M. (1986). Explaining equivocation. *Metaphilosophy*, 17, 192–199.
- Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over and under confidence: The role of error in judgement processes. *Psychological Review*, 101, 519–527.
- Erduan, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education*, 88, 915–933.
- Evans, J. St. B. T., & Over, D. E. (2004). *If*. Oxford, England: Oxford University Press.
- Evans, J. St. B. T., Neilens, H., Handley, S., & Over, D. (2008). When can we say 'if'? *Cognition*, 108, 100–116.
- Fishbein, M., & Ajzen, I. (1981). Acceptance, yielding, and impact: Cognitive processes in persuasion. In R. Petty, T. Ostrom, & T. Brock (Eds.), *Cognitive responses in persuasion* (pp. 339–359). Hillsdale, NJ: Erlbaum.
- Fishbein, M., & Ajzen, I. (2009). *Predicting and changing behavior: The reasoned action approach*. New York: Taylor Francis.
- Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, 90, 239–260.
- Fox, J., Glasspool, D., Grecu, D., Modgil, S., South, M., & Patkar, V. (2007). Argumentation-based inference and decision-making. *IEEE Intelligent Systems*, 22, 34–41.
- Genishi, C., & DiPaolo, M. (1982). Learning through argument in a preschool. In L. C. Wilkinson (Ed.), *Communicating in the classroom* (pp. 49–68). New York: Academic Press.
- Giner-Sorolla, R., Chaiken, S., & Lutz, S. (2002). Validity beliefs and general ideology can influence legal case judgments differently. *Law and Human Behavior*, 26, 507–526.
- Glassner, A., Weinstock, M., & Neuman, Y. (2005). Pupils' evaluation and generation of evidence and explanation in argumentation. *British Journal of Educational Psychology*, 75, 105–118.
- Godden, D. M., & Walton, D. (2004). Denying the antecedent as a legitimate argumentative strategy: A dialectical model. *Informal Logic*, 24, 219–243.
- Goldman, A. I. (1994) Argumentation and social epistemology. *The Journal of Philosophy*, 91, 27–49.
- Hahn, U. (2011). The problem of circularity in evidence, argument and explanation. *Perspectives on Psychological Science*, 6, 172–182.
- Hahn, U., Harris, A. J. L., & Corner, A. J. (2009). Argument content and argument source: An exploration. *Informal Logic*, 29, 337–367.
- Hahn, U., & Oaksford, M. (2006a). A Bayesian approach to informal argument fallacies. *Synthese*, 152, 207–236.
- Hahn, U., & Oaksford, M. (2006b). Why a normative theory of argument strength and why might one want it to be Bayesian? *Informal Logic*, 26, 1–24.
- Hahn, U., & Oaksford, M. (2007a). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review*, 114, 704–732.
- Hahn, U., & Oaksford, M. (2007b). The burden of proof and its role in argumentation. *Argumentation*, 21, 39–61.
- Hahn, U., & Oaksford, M. (2008) Inference from absence in language and thought. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind* (pp. 121–142). New York: Oxford University Press.
- Hahn, U., Oaksford, M., & Bayindir, H. (2005). How convinced should we be by negative evidence? In B. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 887–892). Mahwah, NJ: Erlbaum.
- Hamblin, C. L. (1970). *Fallacies*. London: Methuen.
- Hample, D. (1977). Testing a model of value argument and evidence. *Communication Monographs*, 44, 106–120.
- Harris, A. J. L., & Hahn, U. (2009) Bayesian rationality in evaluating multiple testimonies: Incorporating the role of coherence. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 35, 1366–1373.
- Harris, A. J., Corner, A. J., & Hahn, U. (2009) "Damned by faint praise": A Bayesian account. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 292–297). Austin, TX: Cognitive Science Society.
- Hardman, D. (2009). *Judgement and decision making*. London: John Wiley & Sons.
- Heyse, T. (1997). Why logic doesn't matter in the (philosophical) study of argumentation. *Argumentation*, 11, 211–224.
- Hoeken, H., & Hustinx, L. (2009). When is statistical evidence superior to anecdotal evidence in supporting probability claims. *Human Communication Research*, 39, 491–510.
- Holyoak, K. J., & Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General*, 128, 3–31.
- Hornikx, J. (2005). A review of experimental research on the relative persuasiveness of anecdotal, statistical, causal, and expert evidence. *Studies in Communication Sciences*, 5, 205–216.
- Hornikx, J. (2007). Is anecdotal evidence more persuasive than statistical evidence? A comment on classic cognitive psychological studies. *Studies in Communication Sciences*, 7, 151–164.
- Howson, C., & Urbach, P. (1993) *Scientific reasoning: The Bayesian approach*. La Salle, IL: Open Court.
- Ivkovic, S. K., & Hans, V. P. (2003). Jurors' evaluations of expert testimony: Judging the messenger and the message. *Law and Social Inquiry*, 28, 441–482.

- Jeffrey, R. (2004). *Subjective probability: The real thing*. Cambridge, England: Cambridge University Press.
- Jimenez-Aleixandre, M. P. (2002). Knowledge producers or knowledge consumers? Argumentation and decision making about environmental management. *International Journal of Science Education*, 24, 1171–1190.
- Jimenez-Aleixandre, M. P., Rodriguez, A. B., & Duschl, R. A. (2000). "Doing the lesson" or "doing science": Argument in High School genetics. *Science Education*, 84, 757–792.
- Johnson, R. H. (2000). *Manifest rationality: A pragmatic theory of argument*. Mahwah, NJ: Erlbaum.
- Johnson, B. T., Maio, G. R., & Smith-McLallen, A. (2005). Communication and attitude change: Causes, processes, and effects. In D. Albarracín, B. T. Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 617–669). Mahwah, NJ: Erlbaum.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237–257.
- Kaye, D. H., & Koehler, D. J. (1991). Can jurors understand probabilistic evidence? *Journal of the Royal Statistical Society A*, 154, 75–81.
- Kirwan, C. (1979) Aristotle and the so-called fallacy of equivocation. *Philosophical Quarterly*, 29, 33–46.
- Klaczynski, P. (2000). Motivated scientific reasoning biases, epistemological biases, and theory polarisation: A two process approach to adolescent cognition. *Child Development*, 71, 1347–1366.
- Korb, K. (2004). Bayesian informal logic and fallacy. *Informal Logic*, 23, 41–70.
- Korpan, C. A., Bisanz, G. L., Bisanz, J., & Henderson, J. M. (1997). Assessing literacy in science: Evaluation of scientific news briefs. *Science Education*, 81, 515–532.
- Kortland, K. (1996). An STS case study about students' decision making on the waste issue. *Science Education*, 80, 673–689.
- Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review*, 96, 674–689.
- Kuhn, D. (1991). *The skills of argument*. Cambridge, England: Cambridge University Press.
- Kuhn, D. (1993). Science as argument: Implications for teaching and learning scientific thinking. *Science Education*, 77, 319–337.
- Kuhn, D. (2001). How do people know? *Psychological Science*, 12, 1–8.
- Kuhn, D., & Udell, W. (2003). The development of argument skills. *Child Development*, 74, 1245–1260.
- Kuhn, D., Cheney, R., & Weinstock, M. (2000). The development of epistemological understanding. *Cognitive Development*, 15, 309–328.
- Kuhn, D., Shaw, V., & Felton, M. (1997). Effects of dyadic interaction on argumentative reasoning. *Cognition and Instruction*, 15, 287–315.
- Lamberts, K. (1995). Categorization under time pressure. *Journal of Experimental Psychology: General*, 124, 161–180.
- Laplace, P. S. (1951). *A philosophical essay on probabilities* (F. W. Truscott & F. L. Emory, Trans.). New York: Dover Publications. (Original work published 1814).
- Lord, C. G., & Taylor, C. A. (2009). Biased assimilation: Effects of assumptions and expectations on the interpretation of new evidence. *Social and Personality Psychology Compass*, 3, 827–841.
- Manktelow, K. (2011). *Reasoning and thinking*. Hove, England: Taylor Francis.
- Manktelow, K. I., & Over, D. E. (1987). Reasoning and rationality. *Mind and Language*, 2, 199–219.
- Maio, G. R., & Haddock, G. G. (2010). *The psychology of attitudes and attitude change*. London: Sage.
- Means, M. L., & Voss, J. F. (1996). Who reasons well? Two studies of informal reasoning among children of different grade, ability, and knowledge levels. *Cognition and Instruction*, 14, 139–179.
- Mercier, H., & Sperber, S. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioural and Brain Sciences*, 34, 57–74.
- McGuire, W. J. (1960a). Cognitive consistency and attitude change. *Journal of Abnormal and Social Psychology*, 60, 345–353.
- McGuire, W. J. (1960b). Direct and indirect persuasive effects of dissonance-producing messages. *Journal of Abnormal and Social Psychology*, 60, 354–358.
- McGuire, W. J. (1960c). A syllogistic analysis of cognitive relationships. In C. L. Hovland & M. J. Rosenberg (Eds.), *Attitude organization and change: An analysis of consistency among attitude components* (pp. 65–111). New Haven, CT: Yale University Press.
- McKenzie, C. R. M., Lee, S. M., & Chen, K. K. (2002). When negative evidence increases confidence: Change in belief after hearing two sides of a dispute. *Journal of Behavioral Decision Making*, 15, 1–18.
- McQuiston-Surrett, D., & Saks, M. J. (2009). The testimony of forensic identification science: What expert witnesses say and what fact finders hear. *Law and Human Behavior*, 33, 436–453.
- Mischos, C. (2003). Cognitive, emotional and verbal response in unfair everyday discourse. *Journal of Language and Social Psychology*, 22(1), 119–131.
- Neuman, Y. (2003). Go ahead, prove that God does not exist! *Learning and Instruction*, 13, 367–380.
- Neuman, Y., Glassner, A., & Weinstock, M. (2004). The effect of a reason's truth-value on the judgment of a fallacious argument. *Acta Psychologica*, 116, 173–184.
- Neuman, Y., Weinstock, M. P., & Glasner, A. (2006). The effect of contextual factors on the judgment of informal reasoning fallacies. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 59(A), 411–425.
- Neuman, Y., & Weitzman, E. (2003). The role of text representation in students' ability to identify fallacious arguments. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 56(A), 849–864.
- Nickerson, R. S. (2008). *Aspects of rationality: Reflections on what it means to be rational and whether we are*. Hove, UK: Psychology Press.
- Norris, S. P., & Phillips, L. M. (1999). How literacy in its fundamental sense is central to scientific literacy. *Science Education*, 87, 224–240.
- Norris, S. P., Phillips, L. M., & Korpan, C. A. (2003). University students' interpretation of media reports of science and its relationship to background knowledge, interest and reading difficulty. *Public Understanding of Science*, 12, 123–145.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115, 39–57.
- Nosofsky, R. M. (1988a). Exemplar-based accounts of the relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 14, 700–708.

- Nosofsky, R. M. (1988b). Similarity, frequency and category representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14, 54–65.
- Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101, 608–631.
- Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford, England: Oxford University Press.
- Oaksford, M., & Chater, N. (2009). Precis of “Bayesian rationality: The probabilistic approach to human reasoning.” *Behavioral and Brain Sciences*, 32, 69–120.
- Oaksford, M., & Chater, N. (Eds.). (2010a). *Cognition and conditionals: Probability and logic in human thinking*. Oxford, England: Oxford University Press.
- Oaksford, M., & Chater, N. (2010b). Conditionals and constraint satisfaction: Reconciling mental models and the probabilistic approach? In M. Oaksford & N. Chater (Eds.), *Cognition and conditionals: Probability and logic in human thinking* (pp. 309–334). Oxford, England: Oxford University Press.
- Oaksford, M., & Chater, N. (2010c). Causation and conditionals in the cognitive science of human reasoning. [Special issue. J. C. Perales, & D. R. Shanks, Eds. *Causal learning beyond causal judgment*]. *Open Psychology Journal*, 3, 105–118.
- Oaksford, M., Chater, N., & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 883–899.
- Oaksford, M., & Hahn, U. (2004). A Bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology*, 58, 75–85.
- Oaksford, M., & Hahn, U. (2007). Induction, deduction and argument strength in human reasoning and argumentation. In A. Feeney & E. Heit (Eds.), *Inductive reasoning* (pp. 269–301). Cambridge, England: Cambridge University Press.
- O’Keefe, D. J. (1995). Argumentation studies and dual-process models of persuasion. In F. H. van Eemeren, R. Grootendorst, J. A. Blair, & C. A. Willard (Eds.), *Proceedings of the Third ISSA Conference on Argumentation. Vol. 1: Perspectives and approaches* (pp. 3–17). Amsterdam, Netherlands: Sic Sat.
- O’Keefe, D. J. (1997a). Standpoint explicitness and persuasive effect: A meta-analytic review of the effects of varying conclusion articulation in persuasive messages. *Argumentation and Advocacy*, 34, 1–12.
- O’Keefe, D. J. (1997b). Justification explicitness and persuasive effect: A meta-analytic review of the effects of varying support articulation in persuasive messages. *Argumentation and Advocacy*, 35, 61–75.
- O’Keefe, D. J. (1999). How to handle opposing arguments in persuasive messages: A meta-analytic review of the effects of one-sided and two-sided messages. *Communication Yearbook*, 22–209–256.
- O’Keefe, D. J. (2003). The potential conflict between normatively good argumentative practice and persuasive success. In F. H. van Eemeren, J. A. Blair, C. A. Willard, & A. F. Snoeck Henkemans (Eds.), *Anyone who has a view: Theoretical contributions to the study of argumentation* (pp. 309–318). Dordrecht, Netherlands: Kluwer Academic Publishers.
- Patronis, T., Potari, D., & Spiliotopoulou, V. (1999). Students’ argumentation in decision-making on a socio-scientific issue: Implications for teaching. *International Journal of Science Education*, 21, 745–754.
- Pennington, N., & Hastie, R. (1981). Juror decision-making models: The generalization gap. *Psychological Bulletin*, 89, 246–287.
- Pennington, N., & Hastie, R. (1986). Evidence evaluation in complex decision making. *Journal of Personality and Social Psychology*, 51, 242–258.
- Pennington, N., & Hastie, R. (1988). Explanation-based decision making: Effects of memory structure on judgment. *Journal of Experimental Psychology*, 14, 521–533.
- Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the story model for juror decision making. *Journal of Personality and Social Psychology*, 62(2), 189–206.
- Perelman, C., & Olbrechts-Tyteca, L. (1969). *The new rhetoric: A treatise on argumentation*. Notre Dame, IN: University of Notre Dame Press.
- Perham, N. R., & Oaksford, M. (2005). Deontic reasoning with emotional content: Evolutionary psychology or decision theory? *Cognitive Science*, 29, 681–718.
- Prakken, H. (2008). AI & law on legal argument: Research trends and application prospects. *SCRIPTed*, 5, 449–454. doi: 10.2966/scrif.050308.449.
- Prakken, H., & Vreeswijk, G. A. W. (2002). Logics for defeasible argumentation. In D. M. Gabbay & F. Guenther (Eds.), *Handbook of philosophical logic* (2nd ed., Vol 4, pp. 219–318). Dordrecht/Boston/London: Kluwer Academic Publishers.
- Pyszczynski, T., & Wrightsman, L. S. (1981). The effects of opening statements on mock jurors’ verdicts in a simulated criminal trial. *Journal of Applied Social Psychology*, 11, 301–313.
- Pyszczynski, T., Greenberg, J., Mack, D., & Wrightsman, L. (1981). Opening statements in a jury trial: The effect of promising more than the evidence can show. *Journal of Applied Social Psychology*, 11, 434–444.
- Rahwan, I., & Moraitis, P. (Eds.). (2009). Argumentation in multi-agent systems. Fifth International Workshop, ArgMAS 2008. In *Lecture Notes in Artificial Intelligence* (Vol. 5384). Heidelberg, Germany: Springer.
- Rahwan, I., Zablith, F., & Reed, C. (2007). Laying the foundations for a world wide argument web. *Artificial Intelligence*, 171, 897–921.
- Ramsey, F. P. (1931). *The foundations of mathematics and other logical essays*. London: Routledge and Kegan Paul.
- Ratcliffe, M. (1999). Evaluation of abilities in interpreting media reports of scientific research. *International Journal of Science Education*, 21, 1085–1099.
- Reed, C., & Rowe, G. (2004). Araucaria: Software for argument analysis, diagramming and representation. *International Journal of Artificial Intelligence Tools*, 13, 961–980.
- Rescher, N. (1977). *Dialectics: A controversy oriented approach to the theory of knowledge*. Albany, NY: SUNY Press.
- Ricco, R. B. (2003). The macrostructure of informal arguments: A proposed model and analysis. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 56(A), 1021–1051.
- Rips, L. J. (1998). Reasoning and conversation. *Psychological Review*, 105, 411–441.
- Rips, L. J. (2002). Circular reasoning. *Cognitive Science*, 26, 767–795.
- Rips, L. J., Brem, S. K., & Bailenson, J. N. (1999). Reasoning dialogues. *Current Directions in Psychological Science*, 8, 172–177.

- Sadler, T. D. (2004). Informal reasoning regarding socioscientific issues: A critical review of research. *Journal of Research in Science Teaching*, 41, 513–536.
- Schklar, J., & Diamond, S. S. (1999). Juror reactions to DNA evidence: Errors and expectancies. *Law and Human Behavior*, 23, 159–184.
- Shogenji, T. (2000). Self-dependent justification without circularity. *British Journal for the Philosophy of Science*, 51, 287–298.
- Schreier, M., Groeben, N., & Christmann, U. (1995). "That's not fair!" Argumentational integrity as an ethics of argumentative communication. *Argumentation*, 9, 267–289.
- Schum, D. A. (1981). Sorting out the effects of witness sensitivity and response-criterion placement upon the inferential value of testimonial evidence. *Organizational Behavior and Human Performance*, 27, 153–196.
- Schum, D. A. (1993). Argument structuring and evidence evaluation. In R. Hastie (Ed.), *Inside the juror: The psychology of juror decision making* (pp. 175–191). Cambridge, England: Cambridge University Press.
- Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning*. Evanston, IL: Northwestern University Press.
- Skolnick, P., & Shaw, J. I. (2001). A comparison of eyewitness and physical evidence on mock-juror decision making. *Criminal Justice and Behavior*, 28, 614–630.
- Simon, D., Pham, L. B., Le, Q. A., & Holyoak, K. J. (2001). The emergence of coherence over the course of decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 1250–1260.
- Spellman, B. A., & Tenney, J. R. (2010). Credible testimony in and out of court. *Psychonomic Bulletin and Review*, 17, 168–173.
- Specker, S. C., & Worthington, D. L. (2003). The influence of opening statement/closing argument organizational strategy on juror verdict and damage awards. *Law and Human Behavior*, 27, 437–456.
- Takao, A.Y., & Kelly, G. J. (2003). Assessment of evidence in university students' scientific writing. *Science and Education*, 12, 341–363.
- Thompson, V. A., Evans J. St. B. T., & Handley, S. J. (2005). Persuading and dissuading by conditional argument. *Journal of Memory and Language*, 53, 238–257.
- Tillers, P., & Green, E. (Eds.). (1988). *Probability and inference in the law of evidence: The uses and limits of Bayesianism*. Dordrecht, Netherlands: Kluwer Academic Publishers.
- Tindale, C. W. (2007). *Fallacies and argument appraisal*. New York: Cambridge University Press.
- Toulmin, S. E. (1958). *The uses of argument*. Cambridge, England: Cambridge University Press.
- van Eemeren, F. H., Garssen, B., & Meuffels, B. (2009). *Fallacies and judgments of reasonableness: Empirical research concerning pragmialectical discussion rules*. Dordrecht, Netherlands: Springer.
- van Eemeren, F. H., & Grootendorst, R. (1984). *Speech acts in argumentative discussions. A theoretical model for the analysis of discussions directed towards solving conflicts of opinion*. Berlin, Germany: De Gruyter.
- van Eemeren, F. H., & Grootendorst, R. (1992). *Argumentation, communication, and fallacies*. Hillsdale, NJ: Erlbaum.
- van Eemeren, F. H., & Grootendorst, R. (1987). Fallacies in pragma-dialectical perspective. *Argumentation*, 1, 283–301.
- van Eemeren, F. H., & Grootendorst, R. (2004). *A systematic theory of argumentation. The pragma-dialectical approach*. Cambridge, England: Cambridge University Press.
- van Eemeren, F. H., Grootendorst, R., & Snoeck Henkemans, F. (1996). *Fundamentals of argumentation theory*. Mahwah, NJ: Erlbaum.
- Voss, J. F., & Van Dyke, J. A. (2001). Narrative structure, information certainty, emotional, content, and gender as factors in a pseudo jury decision-making task. *Discourse Processes*, 32, 215–243.
- Walton, D. N. (1980). Why is the ad populum a fallacy? *Philosophy and Rhetoric*, 13, 264–278.
- Walton, D. N. (1985). Are circular arguments necessarily vicious? *American Philosophical Quarterly*, 22, 263–274.
- Walton, D. N. (1987). The ad hominem argument as an informal fallacy. *Argumentation*, 1, 317–331.
- Walton, D. N. (1988). The burden of proof. *Argumentation*, 2, 233–254.
- Walton, D. N. (1991). *Begging the question: Circular reasoning as a tactic in argumentation*. New York: Greenwood Press.
- Walton, D. N. (1992a). Nonfallacious arguments from ignorance. *American Philosophical Quarterly*, 29, 381–387.
- Walton, D. N. (1992b). *Slippery slope arguments*. Oxford, England: Oxford University Press.
- Walton, D. N. (1995). *A pragmatic theory of fallacy*. Tuscaloosa: The University of Alabama Press.
- Walton, D. N. (1996a). *Arguments from ignorance*. Philadelphia: Pennsylvania State University Press.
- Walton, D. N. (1996b). *Fallacies arising from ambiguity*. Dordrecht, Netherlands: Kluwer Academic Publishers.
- Walton, D. N. (1997). *Appeal to expert opinion: Arguments from authority*. University Park: Pennsylvania State University Press.
- Walton, D. N. (1998a). *The new dialectic: Conversational contexts of argument*. Toronto, ON: University of Toronto Press.
- Walton, D. N. (1998b). *Ad hominem arguments*. Tuscaloosa: University of Alabama Press.
- Walton, D. N. (1999). *Appeal to popular opinion*. University Park: Pennsylvania State University Press.
- Walton, D. N. (2000). Case study of the use of a circumstantial ad hominem in political argumentation. *Philosophy and Rhetoric*, 33, 101–115.
- Walton, D. N. (2005). Begging the question in arguments based on testimony. *Argumentation*, 19, 85–113.
- Walton, D. N. (2006a). *Fundamentals of critical argumentation*. Cambridge, England: Cambridge University Press.
- Walton, D. N. (2006b). *Scare tactics: Arguments that appeal to fear and threats* (Argumentation Library Series). Dordrecht, Netherlands: Kluwer Academic Publishers.
- Walton, D. N., & Macagno, F. (2007). The fallaciousness of threats: Character and ad baculum. *Argumentation*, 21, 63–81.
- Walton, D. N., Reed, C., & Macagno, F. (2008). *Argumentation schemes*. Cambridge, England: Cambridge University Press.
- Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content*. Cambridge, MA: Harvard University Press.
- Weinstock, M. P., & Flaton, R. A. (2004). Evidence coverage and argument skills: Cognitive factors in a juror's verdict choice. *Journal of Behavioral Decision Making*, 17, 191–212.

- Whately, R. (1828/1963). *Elements of rhetoric* (D. Ehninger, Ed.). Carbondale: University of Southern Illinois Press.
- Woods, J., & Walton, D. (1979). Equivocation and practical logic. *Ratio*, 21, 31–43.
- Woods, J., Irvine, A., & Walton, D. N. (2004). *Argument: Critical thinking, logic and the fallacies* (Rev. ed.). Toronto, ON: Prentice Hall.
- Wyer, R. S., Jr. (1970). Quantitative prediction of belief and opinion change: A further test of a subjective probability model. *Journal of Personality and Social Psychology*, 16, 559–570.
- Wyer, R. S., Jr., & Goldberg, L. (1970). A probabilistic analysis of the relationships among beliefs and attitudes. *Psychological Review*, 77, 100–120.

PART 3

Judgment and
Decision Making

This page intentionally left blank

Decision Making

Robyn A. LeBoeuf *and* Eldar Shafir

Abstract

This chapter reviews selected psychological research on human decision making. The classical, rational theory of choice holds that decisions reflect consistent, stable preferences, which are unaffected by logically immaterial changes in context, presentation, or description. In contrast, empirical research has found preferences to be sensitive to logically irrelevant changes in the context of decision, in how options are described, and in how preferences are elicited. Decisions are also swayed by affect and by decisional conflict and are often driven by the reasons that are most accessible at the moment of choice, leading to preference reversals when, for example, different reasons are made accessible. More broadly, decision makers tend to adopt a “local” perspective: They accept decisions as described and focus on the most salient attributes, even when a more “global” perspective, less influenced by local context and frame, might yield decisions that are less biased by temporary and irrelevant concerns. Future directions and implications for theory and practice are discussed.

Key Words: choice, uncertainty, loss aversion, framing, preference reversals, intertemporal choice, priming

Introduction

People make countless decisions every day, ranging from ones that are barely noticed and soon forgotten (“What should I drink with lunch?” “What should I watch on TV?”), to others that are highly consequential (“How should I invest my retirement funds?” “Should I marry this person?”). In addition to having practical significance, decision making plays a central role in many academic disciplines: Virtually all of the social sciences—including psychology, sociology, economics, political science, and law—rely on models of decision-making behavior. This combination of practical and scholarly factors has motivated great interest in how decisions are and should be made. Although decisions can differ dramatically in scope and content, research has uncovered systematic regularities in how people make decisions and has led to the formulation of

general psychological principles that characterize decision-making behavior. This chapter provides a selective review of those regularities and principles. (For further reviews and edited collections, see, among others, Bazerman & Moore, 2008; Connelly, Arkes, & Hammond, 2000; Dawes & Hastie, 2001; Goldstein & Hogarth, 1997; Kahneman & Tversky, 2000; Koehler & Harvey, 2004; Lichtenstein & Slovic, 2006; and Weber & Johnson, 2009.)

The classical treatment of decision making, known as the “rational theory of choice” or the “standard economic model,” posits that people have orderly preferences that obey a few simple and intuitive axioms. When faced with a choice, people are assumed to gauge each alternative’s “subjective utility” and to choose the alternative with the highest. In the face of uncertainty about whether outcomes will obtain, people are thought to calculate an option’s subjective

expected utility, which is the sum of its subjective utilities over the possible outcomes weighted by these outcomes' estimated probabilities of occurrence. Deciding then is simply a matter of choosing the option with the greatest expected utility; choice is thus thought to *reveal* a person's subjective utility functions and, hence, her underlying preferences (e.g., Keeney & Raiffa, 1976; Savage, 1954; von Neumann & Morgenstern, 1944).

While highly compelling in principle, the standard view has met with persistent critiques addressing its inadequacy as a description of how decisions are actually made. For example, Simon (1955) suggested replacing the rational model with a framework that accounted for a variety of human resource constraints, such as bounded attention and memory capacity, as well as limited time. According to the bounded rationality view, it was unreasonable to expect decision makers to exhaustively compute options' expected utilities.

Other critiques have focused on systematic violations of even the most fundamental requirements of the rational theory of choice. According to the theory, for example, preferences should remain unaffected by logically inconsequential factors such as the specific procedure used to elicit preferences or the precise manner in which options are described (Arrow, 1951, 1988; Tversky & Kahneman, 1986). However, a series of compelling demonstrations showed that choices failed to obey simple consistency requirements (see Chater & Oaksford, Chapter 2) and were, instead, affected by nuances of the decision context that were not subsumed by the normative accounts (e.g., Lichtenstein & Slovic, 1971, 1973; Tversky & Kahneman, 1981). This research suggested that preferences are constructed, not merely revealed, in the making of decisions, leading to significant and systematic departures from normative predictions (Lichtenstein & Slovic, 2006; Payne, Bettman, & Johnson, 1992; Slovic, 1995).

The mounting evidence has forced a clear division between normative and descriptive treatments. The rational model remains the normative standard against which decisions are often judged, both by experts and by novices (cf. Stanovich, 1999). At the same time, substantial research has made considerable progress in developing models of choice that are descriptively more faithful. Descriptive accounts as elegant and comprehensive as the normative model are not yet (and may never be) available, but research has uncovered robust principles that play a central role in decision making. In what follows,

we review some of these principles, and we consider the fundamental ways in which they conflict with normative expectations.

Choice Under Uncertainty

For some decisions, the availability of options is essentially certain (as when choosing items from a menu, or cars at a dealer's lot). Other decisions are made under uncertainty. They are "risky" when the probabilities of the outcomes are known (e.g., gambling or insurance), or, as with most real-world decisions, they are "ambiguous," in that precise likelihoods are not known and must be estimated by the decision maker. When deciding under uncertainty, a person must consider both the desirability of the potential outcomes and their likelihoods; much research has addressed the manner in which these factors are estimated and combined.

Prospect Theory

When facing a choice between a risky prospect that offers a 50% chance to win \$200 (and a 50% chance to win nothing) versus an alternative of receiving \$100 for sure, most people prefer the sure gain over the gamble, although the two prospects have the same expected value. (The expected value is the sum of possible outcomes with each outcome weighted by its probability of occurrence. The expected value of the gamble above is $.50 * \$200 + .50 * 0 = \100 .) Such preference for a sure outcome over a risky prospect of equal expected value is called *risk aversion*; people tend to be risk averse when choosing between prospects with positive outcomes. The tendency toward risk aversion can be explained by the notion of diminishing sensitivity, first formalized by Daniel Bernoulli (1738/1954). Bernoulli proposed that preferences are better described by *expected utility* than by *expected value*, and he suggested that "the utility resulting from any small increase in wealth will be inversely proportionate to the quantity of goods previously possessed" (p. 25), thus effectively predicting a concave utility function. (A function is concave if a line joining two points on the curve lies below the curve.) The expected *utility* of a gamble offering a 50% chance to win \$200 (and 50% nothing) is $.50 * u(\$200)$, where u is the person's utility function ($u(0) = 0$). As illustrated in Figure 16.1, diminishing sensitivity and a concave utility function imply that the subjective value attached to a gain of \$100 is more than half of the value attached to a gain of \$200: $u(100) > .5 * u(200)$. This entails preference

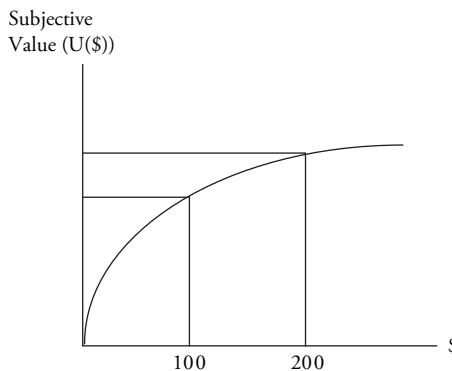


Fig. 16.1 A concave function for gains.

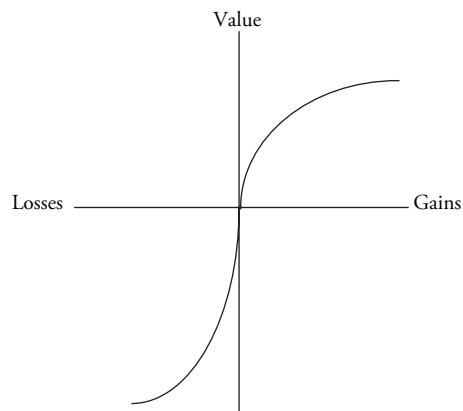


Fig. 16.2 Prospect theory's value function.

for the sure \$100 gain over the gamble and, hence, risk aversion.

However, when asked to choose between a prospect that offers a 50% chance to lose \$200 (and a 50% chance of nothing) versus losing \$100 for sure, most people prefer the risky gamble over the certain loss. This is because diminishing sensitivity applies to negative as well as to positive outcomes: The impact of an initial \$100 loss is greater than that of an additional \$100 loss, implying a convex value function for losses. A gamble offering a 50% chance to lose \$200 thus has a greater (i.e., less negative) expected subjective value than does a sure \$100 loss: $.5^*u(-\$200) > u(-\$100)$, because one does not lose twice as much utility when losing \$200 as when losing \$100. Such preference for a risky prospect over a sure outcome of equal expected value is described as *risk seeking*. With the exception of prospects that involve very small probabilities, risk aversion is generally observed in choices involving gains, whereas risk seeking tends to hold in choices involving losses.

These insights led to the S-shaped value function that forms the basis for prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992), a highly influential descriptive theory of choice. The value function of prospect theory, illustrated in Figure 16.2, has three important properties: (a) it is defined on gains and losses rather than total wealth, capturing the fact that people normally treat outcomes as departures from a current reference point (rather than in terms of final assets, as posited by the rational theory of choice); (b) it is steeper for losses than for gains—thus, a loss of \$X is more aversive than a gain of \$X is attractive, capturing the phenomenon of *loss aversion*; and (c) it is concave for gains and convex for losses, predicting, as described

earlier, risk aversion in the domain of gains and risk seeking in the domain of losses.

In addition, according to prospect theory, probabilities are not treated linearly; instead, people tend to overweight small probabilities and to underweight large ones (Gonzalez & Wu, 1999; Kahneman & Tversky, 1979; Prelec, 2000). Among other things, this nonlinearity has implications for the attractiveness of gambling and insurance (which typically involve low-probability events); and it yields substantial discontinuities at the endpoints, where the passage from impossibility to possibility and from high likelihood to certainty can have inordinate impact (Camerer, 1992; Kahneman & Tversky, 1979). The weighting of probabilities can also be influenced by factors such as the decision maker's feeling of competence in a domain (Heath & Tversky, 1991) or by whether the options involve nonmonetary instead of monetary outcomes (McGraw, Shafir, & Todorov, 2010; Rottenstreich & Hsee, 2001). Such attitudes toward value and chance lead to substantial sensitivity to contextual factors when making decisions, as discussed later in this chapter.

The Framing of Risky Decisions

The previously discussed attitudes toward risky decisions appear relatively straightforward, and yet they yield choice patterns that conflict with normative standards. Perhaps the most fundamental of these patterns is the “framing effect” (Tversky & Kahneman, 1981, 1986). Because risk attitudes differ when outcomes involve gains as opposed to losses, the same decision can elicit conflicting risk attitudes depending on whether it is framed as involving gains or losses. In one example, financial services professionals considered advising a client who had \$6,000

invested in the stock market during a market downturn. Some respondents chose between a strategy that would save \$2,000 of the money with certainty and another strategy that offered a 1/3 chance that all \$6,000 would be saved (but a 2/3 chance that nothing would be saved). Others chose between a strategy in which \$4,000 would be lost with certainty and another that offered a 1/3 chance that no money would be lost (and a 2/3 chance that all \$6,000 would be lost). All respondents thus faced the same choice: Both groups chose between saving \$2,000 (and losing \$4,000) with certainty versus risking all of the money (at 2/3 chance) for a 1/3 chance that nothing will be lost (Roszkowski & Snelbaker, 1990; see Tversky & Kahneman, 1981). People, however, tend to “accept” the provided frame and to consider problems as presented, failing to reframe them from alternate perspectives. As a result, most people considering the problem in terms of money saved (i.e., a choice between “gains”) showed a risk-averse preference for the certain outcome, whereas most of those considering the problem in terms of money lost (i.e., a choice between “losses”) showed a risk-seeking preference for the probabilistic strategy. This pattern violates the normative requirement of “description invariance,” according to which logically equivalent descriptions of a decision should yield the same preference (see Keren, 2011; Kühberger, 1995; Levin, Schneider, & Gaeth, 1998 for reviews).

The acceptance of problem frame, combined with the nonlinear weighting of probabilities and, in particular, with the elevated impact of perceived “certainty,” has a variety of normatively troubling consequences. Consider, for example, the following choice between gambles (Tversky & Kahneman, 1981, p. 455):

- A. A 25% chance to win \$30
- B. A 20% chance to win \$45

Faced with this choice, the majority (58%) of participants preferred option B. Now, consider the following extensionally equivalent problem:

In the first stage of this game, there is a 75% chance to end the game without winning anything, and a 25% chance to move into the second stage. If you reach the second stage, you have a choice between:

- C. A sure win of \$30
- D. An 80% chance to win \$45

The majority (78%) of participants now preferred option C over option D, even though, when

combined with the first stage of the problem, options C and D are equivalent to A and B, respectively. Majority preference thus reverses as a function of a logically irrelevant contextual variation. In this case, the reversal is due to the impact of apparent certainty (which renders option C more attractive than option A) and to people’s tendency to contemplate decisions from a “local” rather than a “global” perspective. As noted, a combination of the two stages in the last problem would have easily yielded the same representation as that of the first problem. But rather than amalgamating across events and decisions, as is often assumed in normative analyses, people tend to contemplate each decision separately, which can yield conflicting attitudes across choices. We return to the issue of local versus global perspectives in a later section.

As a further example of framing, risk attitudes can reverse even within the domain of losses, depending on the context of decision. Thus, people tend to prefer a sure loss to a risky prospect when the sure loss is described as “insurance” against a low-probability, high-stakes loss (Hershey & Schoemaker, 1980). The insurance context brings to mind a social norm, making the insurance premium appear more like an investment than a loss, with the low-probability, high-stakes loss acquiring the character of a neglected responsibility rather than a considered risk (e.g., Hershey & Schoemaker, 1980; Kahneman & Tversky, 1979; Kunreuther & Pauly, 2005; Slovic, Fischhoff, & Lichtenstein, 1988).

The framing of certainty and risk also affects people’s thinking about financial transactions through inflationary times, as illustrated by the following example. Participants were asked to imagine that they were in charge of buying computers (currently priced at \$1,000) that would be delivered and paid for 1 year later, by which time, due to inflation, prices were expected to be approximately 20% higher (and equally likely to be above or below the projected 20%). All participants essentially faced the same choice: They could agree to pay either \$1,200 (20% more than the current price) upon delivery next year, or they could agree to pay the going market price in 1 year, which would depend on inflation. Reference points were manipulated to make one option appear certain while the other appeared risky: Half the participants saw the contracts framed in nominal terms, so that the \$1,200 price appeared certain, whereas the future nominal market price (which could be more or less than \$1,200) appeared risky. Other participants saw the contracts framed in

real terms, so that the future market price appeared appropriately indexed, whereas precommitting to a \$1,200 price, which could be lower or higher than the actual future market price, seemed risky. In both conditions, respondents preferred the contract that appeared certain, preferring the fixed price in the nominal frame, and the indexed price in the “real” frame (Shafir, Diamond, & Tversky, 1997). As with many psychological tendencies, the preference for certainty can mislead in some circumstances, but it may also be exploited for beneficial ends, such as when the certainty associated with a particular settlement is highlighted to boost the chance for conflict resolution (Kahneman & Tversky, 1995).

Riskless Choice

Not all decisions involve risk or uncertainty. For example, when choosing items in a store, we can be fairly confident that the displayed items are available. (Naturally, there could be uncertainty about one’s eventual satisfaction with an item, but we leave those considerations aside for the moment.) The absence of uncertainty, however, does not eliminate preference malleability, and many of the principles discussed earlier affect even riskless decisions. Recall that outcomes can be framed as gains or losses relative to a reference point, that losses typically “loom larger” than comparable gains, and that people tend to accept the presented frame. These factors, even in the absence of risk, can yield problematic decision patterns.

Loss Aversion and the Status Quo

A fundamental fact about the making of decisions is loss aversion: According to loss aversion, the pain associated with giving up a good is greater than the pleasure associated with obtaining it (Tversky & Kahneman, 1991). This yields “endowment effects,” wherein merely possessing a good (such that parting with it is rendered a loss) can lead to higher valuation of the good than if it were not in one’s possession. A classic experiment illustrates this point (Kahneman, Knetsch, & Thaler, 1990). Participants were arbitrarily assigned to be *sellers* or *choosers*. The sellers were each given an attractive mug, which they could keep, and were asked to indicate the lowest amount for which they would sell the mug. The *choosers* were not given a mug but were instead asked to indicate the amount of money that the mug was worth to them. An official market price, \$X, was to be revealed; all those who valued the mug at more than \$X received a mug, whereas

those who valued the mug below \$X received \$X. All participants, whether sellers or choosers, essentially faced the same task of determining a price at which they would prefer money over the mug. Since participants were randomly assigned to be sellers or choosers, standard expectations are that the two groups would value the mugs similarly. Loss aversion, however, suggests that the sellers would set a higher price (for what they were about to “lose”) than the choosers. Indeed, sellers’ median asking price was twice that of choosers.

Another manifestation of loss aversion is a general reluctance to trade, illustrated in a study in which half of the participants were given a decorated mug while the others were given a bar of Swiss chocolate (Knetsch, 1989). Later, each subject was shown the alternative gift and offered the opportunity to trade his or her gift for the other. Because the initial allocation of gifts was arbitrary and transaction costs minimal, economic theory (specifically, the Coase theorem) predicts that about half the participants would exchange their gifts. Loss aversion, on the other hand, predicts that most participants would be reluctant to give up a gift in their possession (a loss) to obtain the other (a gain). Indeed, only 10% of the participants chose to trade (see also Kahneman, Knetsch, & Thaler, 1990). This outcome contrasts sharply with the standard analysis, in which the value of a good does not change when it becomes part of one’s endowment.

Loss aversion thus promotes stability rather than change. In particular, it predicts a strong tendency to maintain the status quo, because the disadvantages of departing from it loom larger than the advantages of its alternative (Samuelson & Zeckhauser, 1988). Consider, for example, two candidates, Frank and Carl, who are running for election during difficult times and have announced target inflation and unemployment figures. Frank proposes a 42% yearly inflation rate and 15% unemployment, whereas Carl envisions 23% inflation and 22% unemployment. When Carl’s figures represent the status quo, Frank’s plans entail greater inflation and diminished unemployment, whereas when Frank’s figures are the status quo, Carl’s plan entails lower inflation and greater unemployment. As predicted, neither departure from the “current” state was endorsed by the majority of respondents, who preferred whichever candidate was said to represent the status quo (Quattrone & Tversky, 1988).

A striking tendency to maintain the status quo was observed in the context of retirement-savings

decisions made by employees at a Fortune 500 company. Employee enrollment in a 401(k) savings plan was compared under two conditions: when employees had to actively take steps to enroll (and thus, nonparticipation was the default) versus when employees were automatically enrolled upon hiring but could take steps to unenroll (and thus participation was the default). In both cases, the transaction costs of switching away from the default were minimal, but the effect of changing the default was substantial: When nonparticipation was the default, only 37% of those who had been employed for between 3 and 15 months had enrolled, but when participation was the default, 86% of those employed for an identical length of time remained enrolled (Madrian & Shea, 2001; see Johnson, Hershey, Meszaros, & Kunreuther, 1993 for a similar example involving automobile insurance). Another naturally occurring experiment was observed in decisions regarding organ donation (Johnson & Goldstein, 2003). In some European nations, people are by default organ donors unless they elect not to be, whereas in other European nations they are, by default, not donors unless they choose to be. Observed rates of organ donors are almost 98% in the former nations and about 15% in the latter, a remarkable difference given the low transaction costs and the significance of the decision.

The status quo bias can affect decisions in many domains (see also Tversky & Kahneman, 1991) and can also hinder the negotiated resolution of disputes. If each disputant sees the opponent's concessions as gains but its own concessions as losses, agreement will be hard to reach because each will perceive itself as relinquishing more than it stands to gain. Because loss aversion renders forgone gains more palatable than comparable losses (cf. Kahneman, 1992), an insightful mediator may do best to set all sides' reference points low, thus requiring compromises over outcomes that are mostly perceived as gains.

Loss aversion continues to inspire research, with recent work identifying potential moderators and boundary conditions. For example, research suggests that loss aversion may not emerge for amounts of money that are small (Harinck, Van Dijk, Van Beest, & Mersmann, 2007) or for goods that are inherently intended to be exchanged, such as money that one intends to spend or products that one intends to sell (Novemsky & Kahneman, 2005). Furthermore, the endowment effect may reverse for alternatives that primarily differ on *negative* dimensions (e.g., two jobs that each have a negative feature). Instead

of being reluctant to trade such items, people seem to be especially likely to trade them, in part because "losing" a negative feature (e.g., giving up working on weekends) looms larger than "gaining" a negative feature (e.g., taking on a longer commute; Brenner, Rottenstreich, Sood, & Bilgin, 2007).

Semantic Framing

The tendency to adopt the provided frame can affect riskless choice via "attribute-framing" effects, which alter the perceived quality of items by changing their descriptions (Levin et al., 1998). A package of ground beef, for example, can be described as 75% lean or as 25% fat. Not surprisingly, it tends to be evaluated more favorably under the former description than the latter (Levin, 1987; see also Levin, Schnittjer, & Thee, 1988). Similarly, a community with a 3.7% crime rate tends to be allocated greater police resources than one described as 96.3% "crime free" (Quattrone & Tversky, 1988). Attribute-framing effects are not limited to riskless choice; for example, people are more favorably inclined toward a medical procedure when its chance of success, rather than failure, is highlighted (Levin et al., 1988).

Part of the impact of such semantic factors may be due to spreading activation (Collins & Loftus, 1975) wherein positive words (e.g., "crime free") activate associated positive concepts, and negative words activate negative concepts. The psychophysical properties of numbers also contribute to these effects. A 96.3% "crime free" rate, for example, appears insubstantially different from 100% and suggests that "virtually all" are law abiding. The difference between 0% and 3.7%, in contrast, appears more substantial and suggests the need for intervention (Quattrone & Tversky, 1988; see Keren, 2011, for more on the mechanisms that contribute to framing effects). Like the risk attitudes described earlier, such perceptual effects often seem natural and harmless in their own right, but they can generate preference inconsistencies that appear perplexing, especially given the rather mild and often unavoidable manipulations (after all, things need to be described one way or another) and the trivial computations required to translate from one frame to another.

Conflict and Reasons

Choices can be hard to make. People often approach difficult decisions by looking for a compelling rationale to choose one option over another.

At times, compelling rationales are easy to come by and to articulate, whereas other times no compelling rationale presents itself, rendering the conflict between options hard to resolve. Such conflict can be aversive and can lead people to postpone the decision or to select a “default” alternative. The tendency to rely on compelling rationales that minimize conflict appears benign; nonetheless, it can generate preference patterns that are fundamentally different from those predicted by normative accounts based on value maximization.

Decisional Conflict

One way to avoid conflict in choice is to opt for what appears like no choice at all, namely, the status quo. In one example (Tversky & Shafir, 1992a), participants who were purportedly looking to buy a CD player were presented with a Sony player that was on a 1-day sale for \$99, well below the list price. Two-thirds of the participants said they would buy such a CD player. Another group was presented with the same Sony player but also with a top-of-the-line Aiwa player selling for \$159. In this group, only 54% expressed interest in buying either option, and a full 46% preferred to wait until they learned more about the various models. The addition of an attractive option increased conflict and diminished the number who ended up with either player, despite the fact that most participants preferred the initial alternative to the status quo. This violates what is known in the normative approach as the regularity condition, according to which the “market share” of an existing option—here, the status quo—cannot be increased by enlarging the offered set (see also Tversky & Simonson, 1993).

A related pattern was observed in a grocery store, where shoppers were offered the chance to taste any of 6 jams in one condition, or any of 24 jams in the other (Iyengar & Lepper, 2000). In the 6-jams condition, 40% of shoppers stopped to have a taste and, of those, 30% proceeded to purchase a jam. In the 24-jam condition, a full 60% stopped to taste, but only 3% purchased. Presumably, the conflict between so many attractive options proved hard to resolve. Further studies found that those choosing goods from a larger set later reported lower satisfaction with their chosen option than those choosing from a smaller set. Conflict among options thus appears to make people less happy about choosing, as well as less happy with their eventual choice.

Decisional conflict tends to favor default alternatives, much as it advantages the status quo. In one

study, students agreed to fill out a questionnaire in return for \$1.50. Following the questionnaire, half of the respondents were offered the opportunity to exchange the \$1.50 (the default) for a choice between two prizes: a metal *Zebra* pen or a pair of plastic *Pilot* pens. The remaining participants were only offered the opportunity to exchange the \$1.50 for the *Zebra*. The pens were shown to participants, who were informed that each prize regularly costs just over \$2.00. Only 25% opted for the payment over the *Zebra* when *Zebra* was the only alternative, but a reliably greater 53% retained the payment when both pen types were offered (Tversky & Shafir, 1992a). Whereas the majority of participants took advantage of the opportunity to obtain a valuable alternative (the *Zebra*) when there was only one such alternative, the availability of competing valuable alternatives increased the tendency to retain the default option.

Related effects have been documented in decisions made by expert physicians and legislators (Redelmeier & Shafir, 1995). In one scenario, neurologists and neurosurgeons were asked to decide which of several patients awaiting surgery ought to be operated on first. Half of the respondents were presented with two patients, a woman in her early fifties and a man in his seventies. Others saw the same two patients along with a third, a woman in her early fifties highly comparable to the first, so that it was difficult to find a rationale for choosing one woman over the other. As predicted, more physicians chose to operate on the older man when the presence of two highly comparable women introduced decisional conflict than when the choice was between only one younger woman and the man (58% versus 38%, respectively).

Thus, the addition of some options can generate conflict and reduce the tendency to choose. Other options, on the other hand, can lower conflict and increase the likelihood of making a choice. *Asymmetric dominance* (or the *attraction effect*) refers to the fact that in a choice between options A and B, a third option, A', can be added that is clearly inferior to A (but not to B), thereby increasing the choice likelihood of A (Huber, Payne, & Puto, 1982). For example, a choice between \$6 and an elegant pen presents some conflict for participants. But when a less attractive pen is added to the choice set, the superior pen clearly dominates the inferior pen. This dominance provides a rationale for choosing the elegant pen and leads to an increase in the percentage of those choosing that pen over the cash. Along related lines, the *compromise effect*

occurs when the addition of a third, extreme option makes a previously available option appear as a reasonable compromise, thus increasing its popularity (Simonson, 1989; Simonson & Tversky, 1992).

Standard normative accounts do not deny conflict, nor, however, do they assume any direct influence of conflict on choice. (For people who maximize utility, there does not appear to be much room for conflict: Either the utility difference is large and the decision is easy, or it is small and the decision is of little import.) In actuality, however, people are concerned with making the “right” choice. This concern can render decisional conflict influential beyond mere considerations of value. Conflict is an integral aspect of decision making, and the phenomenology of conflict, which can be manipulated via the addition or removal of alternatives, yields predictable and systematic violations of standard normative predictions.

Reason-Based Choice

The desire to make the “right” choice often leads people to look for good reasons when making decisions, and considering this reliance on reasons helps make sense of phenomena that appear puzzling from the perspective of value maximization (Shafir, Simonson, & Tversky, 1993). Relying on good reasons seems like sound practice; after all, the converse, making a choice without good reason, seems unwise. At the same time, abiding by this practice can be problematic because the reasons that come to mind often are fleeting, are limited to what is introspectively accessible, and are not necessarily those that guide, or ought to guide, the decision. For example, participants who were asked to analyze *why* they felt the way that they did about a set of jams showed less agreement with “expert” ratings of the jams than did those who did not analyze their preferences (Wilson & Schooler, 1991). Similarly, among participants who chose a poster to take home, those who analyzed their reasons for liking the poster were less satisfied with their choice weeks later, compared to those who did not analyze their reasons (Wilson et al., 1993). A search for reasons can alter preference in line with reasons that come readily to mind, but those reasons may be heavily influenced by the momentary context. A heavy focus on a biased set of temporarily available reasons can cause one to lose sight of one’s (perhaps more valid) initial feelings (Wilson, Dunn, Kraft, & Lisle, 1989).

Furthermore, a wealth of evidence suggests that people are not always aware of their reasons for

acting and deciding (Nisbett & Wilson, 1977). In one example, participants, presented with four identical pairs of stockings and asked to select one, showed a marked preference for the option on the right. However, despite the evidence that choice was governed by position, no participant mentioned position as the reason for the choice. Respondents easily generated “reasons” (in which they cited attributes such as stocking texture), but the reasons they provided bore little resemblance to those that actually guided choice (Nisbett & Wilson, 1977).

Finally, and perhaps most normatively troubling, a reliance on reasons can induce preference inconsistencies because contextual nuances can render certain reasons more or less apparent. In one study (Tversky & Shafir, 1992b), college students were asked to imagine that they had just passed a difficult exam and now had a choice for the Christmas holidays: They could buy an attractive vacation package at a low price, they could forgo the vacation package, or they could pay a \$5 fee to defer the decision by a day. The majority elected to buy the vacation package, and less than a third elected to delay the decision. A second group was asked to imagine that they had failed the exam and would need to retake it after the holidays. They were then presented with the same choice and, as before, the majority elected to buy the vacation package, and less than a third preferred to defer. However, when a third group of participants was asked to imagine that they did not know whether they had passed or failed the exam, the majority preferred to pay to defer the decision until the next day, when the exam result would be known. Only a minority was willing to commit to the trip without knowing. Apparently, participants were comfortable booking the trip when they had clear reasons for the decision—celebrating when they passed the exam or recuperating when they had failed—but were reluctant to commit when their reasons for the trip were uncertain. This pattern, which violates the Sure Thing Principle (Savage, 1954), has been documented in a variety of contexts, including gambling and strategic interactions (e.g., Prisoner’s Dilemmas; see also Shafir, 1994; Shafir & Tversky, 1992).

The tendency to delay decision for the sake of further information can have a significant impact on the ensuing choice. Consider the following scenario (Bastardi & Shafir, 1998):

You are considering registering for a course in your major that has very interesting subject

matter and will not be offered again before you graduate. While the course is reputed to be taught by an excellent professor, you have just discovered that he will be on leave, and that a less popular professor will be teaching the course.

One group was asked whether they would register for the course, and a majority (82%) said they would. Another group was presented with the same scenario but told that they would not know until the next day whether the regular or the less popular professor would be teaching the course. They could wait until the following day (when they would know who will be teaching) to decide whether to take the course; 56% elected to wait. Those who chose to wait then learned that the less popular professor would be teaching the course; upon receiving the news, nearly half of this group decided not to take the course. Overall, a total of 29% chose not to register for the course in the uncertain version (where they pursued the information) as compared to only 18% in the simple version (when the information was known from the start), a significant 60% increase in turning down the course. The decision to pursue information apparently focuses attention on the information obtained and brings forth particular decision rationales, ultimately distorting preference (Bastardi & Shafir, 1998). Similar patterns have been replicated in a variety of contexts, including one involving nurses in a renal-failure ward, more of whom expressed willingness to donate a kidney (to a hypothetical relative) when they had purportedly elected to be tested and then learned that they were eligible donors than when they had known they were eligible from the start (Redelmeier, Shafir, & Aujla, 2001). A reliance on reasons in choice leaves decision makers susceptible to a variety of contextual and procedural nuances that render salient alternative potential reasons, and thus may lead to inconsistent choices.

Processing of Attribute Weights

Choices can be complex, requiring the evaluation of multiattribute options. Consider, for example, a choice between two job candidates: One did well in school but has relatively unimpressive work experience and moderate letters of recommendation, whereas the other has a poor scholastic record but better experience and stronger letters. To make this choice, the decision maker must somehow combine the attribute information, which requires determining not only the quality or value of each attribute but

also the extent to which a shortcoming on one attribute can be compensated for by strength on another.

Attribute evaluation may be biased by a host of factors known to hold sway over human judgment (for a review, see Griffin, Chapter 17). Moreover, researchers have long known that people have limited capacity for combining information across attributes; because of unreliable attribute weights in human judges, simple linear models tend to yield normatively better predictions than the very judges on whom the models are based (Dawes, 1979; Dawes, Faust, & Meehl, 1989). In fact, people's unreliable weighting of attributes makes them susceptible to a host of manipulations that alter attribute weights and yield conflicting preferences (see Shafir & LeBoeuf, 2004, for a further discussion of multiattribute choice).

Compatibility

One way in which attribute weighting can be manipulated is by changing how a preference is elicited. Imagine a person deciding among simple monetary gambles, which differ on payoffs and the chance to win. That person's preferences can be assessed in different, but logically equivalent, ways (see Schkade & Johnson, 1989, for a review). For example, participants may be asked to choose among the gambles or, alternatively, they may estimate their maximum willingness to pay for each gamble. Notably, these logically equivalent preference elicitation procedures often result in differential weightings of attributes and, consequently, in inconsistent preferences.

Consider two gambles: One offers an 8/9 chance to win \$4 and the other a 1/9 chance to win \$40. People typically *choose* the high-probability, low-payoff gamble but assign a higher *price* to the high-payoff, low-probability gamble, thus expressing conflicting preferences (Grether & Plott, 1979; Lichtenstein & Slovic, 1971, 1973; Tversky, Slovic, & Kahneman, 1990). This pattern illustrates the principle of *compatibility*, according to which an attribute's weight is enhanced by its compatibility with the response mode (Slovic, Griffin, & Tversky, 1990; Tversky, Sattath, & Slovic, 1988): A gamble's potential payoff is weighted more heavily in pricing, where the price and the payoff are in the same monetary units, than in choice, where neither attribute maps onto the response scale (Schkade & Johnson, 1989). As a consequence, the high-payoff gamble is valued more in pricing relative to choice.

For another type of response compatibility, imagine having to choose, versus having to reject, one of

two options. Logically speaking, the two tasks are interchangeable: If people prefer one option, they will reject the second, and vice versa. However, people tend to focus on the relative strengths of options (compatible with choosing) when they choose and on weaknesses (compatible with rejecting) when they reject. As a result, options' positive features loom larger in choice, whereas their negative features are weighted relatively more during rejection. In one study, respondents were presented with pairs of options, such as vacation destinations, that included an enriched option, with various positive and negative features (e.g., gorgeous beaches and great sunshine, but cold water and strong winds), and an impoverished option, with no real positive or negative features, hence neutral in all respects (Shafir, 1993). Some respondents were asked which destination they preferred; others decided which to forgo. Because positive features are weighed more heavily in choice and negative features matter relatively more during rejection, the enriched destination was most frequently chosen *and* rejected. Overall, its choice and rejection rates summed to 115%, significantly more than the impoverished destination's 85%, and more than the 100% expected if choice and rejection were complementary (see also Downs & Shafir, 1999; Wedell, 1997), consistent with the notion that attributes are weighted more heavily when compatible with the response mode.

Separate Versus Comparative Evaluation

Decision contexts can also facilitate or hamper attribute evaluation, and this can alter attribute weights. Not surprisingly, an attribute whose value is clear can have greater impact than an attribute whose value is vague. Similarly, attributes may prove difficult to gauge in isolation but easier to evaluate in comparative settings, leading to what are known as *evaluability* effects. In one study, participants considered two second-hand music dictionaries: one with 20,000 entries but a slightly torn cover and another with 10,000 entries and an unblemished cover. Because people have only a vague notion of how many entries to expect in a music dictionary, when participants saw these one at a time, they were willing to pay more for the dictionary with the new cover than for the one with a cover that was slightly torn. When the dictionaries were evaluated concurrently, however, the number-of-entries attribute became easier to evaluate and was weighted more heavily: Most participants preferred the dictionary with more

entries, despite the inferior cover (Hsee, 1996; Hsee, Loewenstein, Blount, & Bazerman, 1999).

As another example, consider a job that pays \$80,000 a year at a firm where one's peers receive \$100,000, compared to a job that pays \$70,000 while coworkers are paid \$50,000. Consistent with the fact that most people prefer higher incomes, a majority of second-year MBA students who compared the two options preferred the job with the higher absolute—despite the lower relative—income. When the jobs are contemplated separately, however, the precise merits of one's own salary are hard to gauge, but earning less than comparable others renders the former job relatively less attractive than the latter, where one's salary exceeds that of one's peers. Indeed, the majority of MBA students who evaluated the two jobs separately anticipated higher satisfaction in the job with the lower salary but the higher relative position, obviously putting more weight on relative pay in the context of separate evaluation (Bazerman, Schroth, Shah, Diekmann, & Tenbrunsel, 1994).

In the same vein, decision principles that are hard to apply in isolated evaluation may prove decisive in comparative settings, producing systematic fluctuations in attribute weights. Kahneman and Ritov (1994), for example, asked participants about their willingness to contribute to several environmental programs. One program was geared toward saving dolphins in the Mediterranean Sea; another funded free medical checkups for farm workers at risk for skin cancer. When asked which program they would rather support, the vast majority chose the medical checkups for farm workers, presumably following the principle that human lives come before those of animals. However, when asked separately for the largest amount they would be willing to pay for each intervention, respondents, moved by the animals' vivid plight, were willing to pay more for the dolphins than for workers' checkups. In a similar application, potential jurors awarded comparable dollar amounts to plaintiffs who had suffered either physical or financial harm, as long as the cases were evaluated separately. However, in concurrent evaluation, award amounts increased dramatically when the harm was physical as opposed to financial, affirming the notion that physical harm is the graver offense (Sunstein, Kahneman, Schkade, & Ritov, 2002).

Attribute weights, which are normatively assumed to remain stable, systematically shift and give rise to patterns of inconsistent preferences. Notably, discrepancies between separate versus concurrent

evaluation have profound implications for intuition and for policy. Outcomes in life are typically experienced one at a time: A person lives through one scenario or another. Normative intuitions, on the other hand, typically arise from concurrent introspection: We entertain a scenario along with its alternatives. When an event triggers reactions that stem from its being experienced in isolation, important aspects of the experience will be misconstrued by intuitions that arise from concurrent evaluation (Shafir, 2002; see also Hsee & Zhang, 2004).

Local Versus Global Perspectives

Many of the inconsistencies described earlier would not have arisen were decisions considered from a more global perspective. The framing of decisions, for instance, would be of little consequence were people to go beyond the provided frame to represent the decision outcomes in a description-independent, canonical manner. Instead, people tend to accept the decision problem as it is presented, largely because they may not have thought of other ways to look at the decision and also because they may not expect their preferences to be susceptible to presumably incidental alterations. (Note, incidentally, that even if they were to recognize the existence of multiple perspectives, people may still not know how to arrive at a preference independent of a specific formulation; cf. Kahneman, 2003.) In this section, we review several additional decision contexts where a limited or myopic approach guides decision making, and where inconsistent preferences arise as a result of a failure to adopt a more “global” perspective. Such a perspective requires one to ignore momentarily salient features of the decision in favor of other, often less salient, considerations that have long-run consequences.

Repeated Decisions

Decisions that occur on a regular basis are often more meaningful when evaluated “in the long run.” For example, the choice to diet or to exercise makes little difference on any one day and can only be carried out under a long-term perspective that trumps the person’s short-term preferences for cake over vegetables, or for sleeping late rather than going to the gym early. People, however, often do not take this long-term perspective when evaluating instances of a recurring choice; instead, they tend to treat each choice as an isolated event.

In one study, participants were offered a 50% chance to win \$2,000 and a 50% chance to lose

\$500. Although most participants refused to play this gamble once, the majority were eager to play the gamble five times, and, when given the choice, preferred to play the gamble six times rather than five. Apparently, fear of possibly losing the single gamble is compensated for by the high likelihood of ending up ahead in the repeated version. Of note, other participants were asked to imagine that they had already played the gamble five times (outcome as yet unknown) and were given the option to play once more. In this formulation, a majority of participants rejected the additional play. Although participants preferred to play the gamble six times rather than five, once they had finished playing five, the additional opportunity was immediately “segregated” and treated as a single instance, which participants preferred to avoid (Redelmeier & Tversky, 1992). The tendency to frame gambles “narrowly” (i.e., isolated from other risks) may help to explain people’s reluctance to invest in the stock market, despite historically high mean returns (Barberis, Huang, & Thaler, 2006). Indeed, the attractiveness of stocks increases when people consider long-term (instead of short-term) return distributions, suggesting they do not (or cannot) spontaneously convert information about isolated outcomes into an aggregate representation (Benartzi & Thaler, 1999).

In a related vein, physicians can think of their patients “individually” (i.e., patient by patient) or “globally” (e.g., as groups of patients with similar problems). Physicians are more likely to take extra measures, such as ordering expensive tests or recommending an in-person consultation, when they consider the treatment of an individual than when they consider a larger group of similarly afflicted patients (Redelmeier & Tversky, 1990). Personal concerns loom larger when patients are considered individually than when “patients in general” are considered, with the latter group more likely to highlight efficiency concerns. Because physicians tend to see patients one at a time, this predicts a pattern of individual decisions that is inconsistent with what these physicians would endorse from a more global perspective. For a more mundane example, people report greater willingness to wear a seatbelt—and to support pro-seatbelt legislation—when they are shown statistics concerning the lifetime risk of being in a fatal accident, instead of the dramatically lower risk associated with any single auto trip (Slovic et al., 1988; see Kahneman, 2003 for a discussion of the tendency to adopt “narrow” framings).

Similar patterns prompted Kahneman and Lovallo (1993) to argue that decision makers often err by treating each decision as unique, rather than categorizing it as one in a series of similar decisions made over a lifetime (or, in the case of corporations, made by many workers). They distinguish an “inside view” of situations and plans, characterized by a focus on the peculiarities of the case at hand, from an “outside view,” guided by an analysis of a large number of similar cases. Whereas an outside view based, for example, on base rates, typically leads to a more accurate evaluation of the current case, people routinely adopt an inside view, which typically overweights the particulars of the given case at the expense of base-rate considerations. Managers, for example, despite knowing that past product launches have routinely run over budget and behind schedule, may convince themselves that this time will be different because the team is excellent or the product exceptional. The inside view can generate overconfidence (Kahneman & Lovallo, 1993), as well as undue optimism regarding the chances of completing projects on time (e.g., the planning fallacy, Buehler, Griffin, & Ross, 1994) or under budget (Peetz & Buehler, 2009). The myopia that emerges from treating repeated decisions as unique leads to overly bold predictions and to the neglect of considerations that ought to matter in the long run.

Mental Accounting

Specific forms of myopia arise in the context of “mental accounting,” the behavioral equivalent of accounting done by firms, wherein people reason about and make decisions concerning matters such as income, spending, and savings. Contrary to the assumption of “fungibility,” according to which money in one account, or from one source, is a perfect substitute for money in another, the labeling of accounts and the nature of transactions have a significant impact on people’s decisions (Thaler, 1999). For example, people’s reported willingness to spend \$25 on a theater ticket is unaffected by having incurred a \$50 parking ticket, but it is significantly lowered when \$50 is spent on a ticket to a sporting event (Heath & Soll, 1996). Respondents apparently bracket expenses into separate accounts, so that spending on entertainment is affected by a previous entertainment expense in a way that it is not if that same expense is “allocated” to, say, travel. Along similar lines, people who had just lost a \$10 bill were happy to buy a \$10 ticket for a play, but were less willing to buy the ticket if, instead of the money, they had just lost a similar \$10 ticket

(Tversky & Kahneman, 1981). Apparently, participants were willing to spend \$10 on a play even after losing \$10 cash, but they found it aversive to spend what was coded as \$20 on a ticket.

Finally, consider the following scenario, which respondents saw in one of two versions:

Imagine that you are about to purchase a jacket for \$125 [\$15] and a calculator for \$15 [\$125]. The calculator salesman informs you that the calculator you want to buy is on sale for \$10 [\$120] at the other branch of the store, located 20 minutes drive away. Would you make the trip to the other store? (Tversky & Kahneman, 1981, p. 457)

Faced with the opportunity to save \$5 on a \$15 calculator, a majority of respondents agreed to make the trip. However, when the calculator sold for \$125, only a minority was willing to make the trip for the same \$5 savings. A global evaluation of either version yields a 20-minute voyage for \$5 savings; people, however, seem to make decisions based on “topical” accounting (Kahneman & Tversky, 1984), wherein the same \$5 saving is coded as a substantial ratio in one case, and as quite negligible in the other.

Specific formulations and contextual details are neither spontaneously reformulated nor translated into more comprehensive or canonical representations. As a consequence, preferences prove highly labile and dependent on what are often theoretically, as well as practically, unimportant and accidental details. An extensive literature on mental accounting, as well as behavioral finance, forms part of the growing field of behavioral economics (see, e.g., Camerer, Loewenstein, & Rabin, 2004; Thaler 1993, 1999).

Decisions Regarding the Future

TEMPORAL DISCOUNTING

A nontrivial task is to decide how much weight to give to outcomes extended into the future. Various forms of uncertainty (regarding nature, one’s own tastes, and so on) justify some degree of discounting in calculating the present value of future goods. Thus, \$1,000 received next year is typically worth less than \$1,000 received today. As it turns out, observed discount rates tend to be unstable and are often influenced by factors, such as the size of the good and its temporal distance, that are not subsumed under standard normative analyses (see Ainslie, 2001; Berns, Laibson, & Loewenstein, 2007; Frederick, Loewenstein, & Donoghue, 2002; Loewenstein & Thaler, 1989, for review). For example, although

some people prefer an apple today over two apples tomorrow, virtually nobody prefers one apple in 30 days over two apples in 31 days (Thaler, 1981). Because discount functions are nonexponential (see also Loewenstein & Prelec, 1992), a 1-day delay has greater impact when that day is near than when it is far. Similarly, when asked what amount of money in the future would be comparable to receiving a specified amount today, people require about \$60 in 1 year to match \$15 now, but they are satisfied with \$4,000 in a year instead of \$3,000 today. This implies discount rates of 300% in the first case and 33% in the second. To the extent that one engages in a variety of transactions throughout time, imposing wildly disparate discount rates on smaller versus larger amounts ignores the fact that numerous small amounts will eventually add up to be larger, yielding systematic inconsistency.

Discount rates fluctuate even when all the particulars of a transaction are held constant and only “irrelevant” aspects change. For example, people exhibit higher discount rates when future transactions are described in terms of the amount of time until their occurrence (“in 3 months”) than in terms of their dates of occurrence (“on June 13th”; LeBoeuf, 2006; Read, Frederick, Orsel, & Rahman, 2005). The amount-of-time frame may focus more attention on the intervening delay and thus make the transaction feel farther away. Indeed, some have argued that an understanding of the perception of future time can shed light on some discounting anomalies. For example, although discount rates increase as time periods shrink (e.g., people discount more per unit time when waiting for 3 months than when waiting for a year; Thaler, 1981), this may happen in part because the perception of time is nonlinear: One year may not seem four times as long as 3 months. When discount rates are calculated with respect to *subjective* time, rates no longer fluctuate with the (subjective) size of the time interval (Zauberman, Kim, Malkoc, & Bettman, 2009). Along similar lines, future losses seem nearer in time than do equidistant future gains (Bilgin & LeBoeuf, 2010), a finding that may partially explain why losses are discounted less steeply than gains (i.e., the “sign effect”; Thaler, 1981).

Excessive discounting yields myopia, which is often observed in people’s attitudes toward future outcomes (see e.g., Elster & Loewenstein, 1992). In one experiment, the high-school dropout rate was reduced by one-third when dropouts were threatened with the loss of their driving privileges

(Loewenstein & Thaler, 1989). This immediate consequence apparently had a significantly greater impact than the far more serious but more distant socioeconomic implications of failing to graduate from high school. Loewenstein and Thaler also discuss physicians’ typical lament that warning about the risk of skin cancer from excessive sun exposure has less effect than the warning that such exposure can cause large pores and acne. In fact, “quit smoking” campaigns have begun to stress the immediate benefits of quitting (e.g., normal heart rate restored within 20 minutes, reduction of carbon monoxide in the blood within 12 hours) even more prominently than the long-term benefits (American Lung Association, 2011). Similar reasoning applies in the context of promoting safe-sex practices and medical self-examinations, where immediate gratification or discomfort often trumps much greater, but temporally distant, considerations.

FUTURE EVENTS

Whereas some of the research reviewed earlier looks at the perceived distance of future events, other research has examined how future events themselves are perceived. Distant future events tend to be represented abstractly, in terms of a few high-level features, whereas more proximal events are represented concretely, in richer low-level detail (Trope & Liberman, 2003). These differing representations can lead to preference reversals, especially when high- and low-level features differ in valence. For example, people may commit to a conference presentation in the distant future (when the abstract benefits of such presentations are salient) but may regret the decision as it becomes nearer and concrete thoughts about the hassle of traveling arise. Changes in construal level can be induced not just by changes in temporal distance but also by incidental tasks that prompt more concrete or abstract thinking, causing preferences to fluctuate even when the temporal distance to an event remains fixed, a finding with implications for self-control and decision making (Fujita, Trope, Liberman, & Levin-Sagi, 2006; Trope & Liberman, 2003; Trope, Liberman, & Wakslak, 2007). Again, people appear not to consider decisions comprehensively and are instead influenced by features that, often arbitrarily, happen to be salient.

Frames of Mind

Inconsistent decisions can also occur when highly transient frames of mind are momentarily triggered,

highlighting values and desires that may not reflect one's more global preferences. These influences may be automatic, making them sometimes difficult to avoid, detect, or correct (see Wegner & Bargh, 1998, for a review of automaticity; for a related analysis of intuitive judgment, see Kahneman, 2003 and Kahneman & Frederick, 2005). Because choices often involve delayed consumption, failure to appreciate the labile nature of preferences may lead people to select currently favored, but later disliked, options.

PRIMING

At the most basic level, transient mindsets arise when specific criteria, desires, or goals are made momentarily accessible. Grocery shopping while very hungry, for example, is likely to lead to purchases that would not be made under normal circumstances (cf. Loewenstein, 1996). In one study of the susceptibility to temporary goal accessibility, participants first completed a task in which they were exposed to words related to either prestige goals or thrift goals. In an ostensibly separate task, participants then chose between more expensive Nike-brand socks or less expensive Hanes-brand socks. Those for whom thrift goals had been primed were more likely to choose Hanes than those for whom prestige goals had been primed (Chartrand, Huber, Shiv, & Tanner, 2008). Momentary priming thus affected ensuing preferences, rendering accessible certain criteria that had not previously been considered important, despite the fact that product consumption was likely to occur long after such momentary criterion salience dissipated (see Bettman & Sujan, 1987; Mandel & Johnson, 2002; Verplanken & Holland, 2002; Wheeler & Berger, 2007).

IDENTITIES

At a broader level, preferences fluctuate along with momentarily salient identities. A working woman, for example, might think of herself primarily as a mother when in the company of her children but may see herself primarily as a professional while at work. The list of potential identities can be extensive (Turner, 1985), with some of a person's identities (e.g., "mother") conjuring up strikingly different values and ideals from others (e.g., "CEO"). Although choices might be expected to reveal stable and coherent preferences that correspond to the wishes of the self as a whole, choice often fluctuates along with happenstance fluctuations in the salience of these identities. In one study,

college students whose "academic" identities had been made salient were more likely to opt for more academic periodicals (e.g., *The Economist*) than were those whose "socialite" identities had been evoked. Similarly, Chinese Americans whose American identities were evoked adopted more stereotypically American preferences (e.g., for individuality and competition over collectivism and cooperation) than when their Chinese identities had been made salient (LeBoeuf, Shafir, & Bayuk, 2010). Preference tends to align with currently salient identities and even with attitudes that are semantically linked to those identities (Morris, Carranza, & Fox, 2008). This creates systematic tension anytime there is a mismatch between the identity that does the choosing and the one that does the consuming. Indeed, people enjoy their chosen options less when the identity salient at the moment of consumption conflicts with, rather than matches, the identity that was salient at the moment of choice (LeBoeuf et al., 2010).

EMOTIONS AND DRIVES

Emotions can have similar effects, influencing the momentary evaluation of outcomes and, thus, choice. The anticipated pain of a loss is apparently greater for people in a positive than in a negative mood; this leads to greater risk aversion among those in a good mood as they strive for "mood maintenance" (e.g., Isen, Nygren, & Ashby, 1988). Furthermore, risk judgments tend to be more pessimistic among people in a negative than in a positive mood (e.g., Johnson & Tversky, 1983).

However, valence is not the sole determinant of an emotion's influence. For example, anger, a negative emotion, seems to increase appraisals of individual control, leading to optimistic risk assessment and to risk seeking, whereas fear, also a negative emotion, is not associated with appraisals of control and promotes risk aversion (Lerner & Keltner, 2001; Lerner & Tiedens, 2006). Disgust, on the other hand, evokes an implicit tendency to "expel" objects from one's possession; thus, people who had just watched a disgusting movie clip set a lower price for an item they were selling than did people in a neutral mood (Lerner, Small, & Loewenstein, 2004). Sadness, on the other hand, evokes a goal to change one's circumstances and to repair them; people who are sad actually show a reversal of the classic endowment effect, as they are (comparatively) eager to part with items they own but also eager to acquire new items (Lerner et al., 2004; see also Cryder, Lerner, Gross, & Dahl, 2008).

Emotions, or “affect,” also influence the associations or images that come to mind in decision making. Because images can be consulted quickly and effortlessly, an “affect heuristic” has been proposed, wherein the affective assessment of options and outcomes guides decisions (Slovic, Finucane, Peters, & MacGregor, 2002). Furthermore, “anticipatory emotions” (e.g., emotional reactions to being in a risky situation) can influence the cognitive appraisal of decision situations and can affect choice (Loewenstein, Weber, Hsee, & Welch, 2001), just as drives and motivations can influence reasoning more generally (see Molden & Higgins, Chapter 20). Emotion and affect influence decision making, but because they are transient, with emotions and drives constantly fluctuating, such influence contributes to reversals of preferences.

Inconsistency thus often arises because people do not realize that their preferences are being momentarily altered by situationally induced sentiments. Evidence suggests, however, that even when people are aware of being in the grip of a transient drive or emotion, they are not always able to “correct” adequately. Instead, people tend to exhibit “empathy gaps,” wherein they underestimate the degree to which various contextual changes will alter their drives, emotions, and preferences (e.g., Van Boven, Dunning, & Loewenstein, 2000). For example, cigarette smokers who had just smoked underestimated the strength of the cravings they would feel following a period of cigarette deprivation. Despite the fact that these smokers were familiar with such craving, they could not simulate its strength when not actually experiencing it, which might help explain why smokers are overly optimistic about the ease of quitting (Sayette, Loewenstein, Griffin, & Black, 2008). Empathy gaps can further contribute to myopic decisions, as people honor present feelings and inclinations, not fully appreciating the extent to which these may be attributable to factors that may soon dissipate.

METACOGNITIVE INFLUENCES

Decisions can also be influenced by metacognitive experiences that arise while processing stimuli. Much recent research has documented the effects that fluency, the subjective experience of ease or difficulty people have when processing information, can have on decisions (see Alter & Oppenheimer, 2009 for a review). For example, when choice options are more difficult to process (e.g., because they are in a hard-to-read font), people are more likely to defer a

decision than when the options are easier to process (e.g., in a clearer font), although the options themselves have not changed (Novemsky, Dhar, Schwarz, & Simonson, 2007). As another example, stocks with more pronounceable names and ticker symbols outperform those with less pronounceable names and symbols on their initial days of trading, apparently because people make more positive inferences about stocks that are more fluently processed (Alter & Oppenheimer, 2006). Ease of processing can thus be seen to inform decision behavior, even when it is triggered by features irrelevant to the decision (see also Simmons & Nelson, 2006).

Conclusions and Future Directions

A review of the behavioral decision-making literature shows people’s preferences to be highly malleable and systematically affected by a host of factors not subsumed under the compelling and popular normative theory of choice. People’s preferences are heavily shaped, among other things, by particular perceptions of risk and value, by influences on attribute weights, by the tendency to avoid decision conflict and to rely on compelling reasons for choice, by salient identities and emotions, and by a general tendency to accept decision situations as they are described, rarely reframing them in alternative, let alone canonical, ways.

It is tempting to attribute many of the effects to shallow processing or to a failure to consider the decision seriously (see, e.g., Grether & Plott, 1979; Smith, 1985; see Shafir & LeBoeuf, 2002, for discussion of such critiques). For example, it seems plausible that people who consider a problem more carefully might notice that it can be framed in alternate ways. This would allow a consideration of the problem from multiple perspectives and perhaps lead to a response unbiased by problem frame or by other “inconsequential” factors (cf. Sieck & Yates, 1997). Evidence suggests, however, that the patterns discussed here cannot be attributed to laziness, inexperience, or lack of motivation. The same general effects are observed when participants are provided with greater incentives (Grether & Plott, 1979; see Camerer & Hogarth, 1999 for a review), when they are asked to justify their choices (Fagley & Miller, 1987; LeBoeuf & Shafir, 2003; Levin & Chapman, 1990), when they are experienced or expert decision makers (Camerer, Babcock, Loewenstein, & Thaler, 1997; McNeil, Pauker, Sox, & Tversky, 1982; Redelmeier & Shafir, 1995; Redelmeier, Shafir, & Aujla, 2001), when they are high in cognitive ability

(Stanovich & West, 2008), or when they naturally think more deeply about problems (LeBoeuf & Shafir, 2003; Levin, Gaeth, Schreiber, & Lauriola, 2002; see Weber & Johnson, 2009 for a review of individual differences in decision making). These findings suggest that many of the attitudes triggered by specific choice frames are at least somewhat entrenched, with extra thought or effort only serving to render the dominant perspective more compelling, rather than highlighting the need for debiasing (Arkes, 1991; LeBoeuf & Shafir, 2003; Thaler, 1991).

Research in decision making is active and growing. Among interesting current developments, a thriving stream of research is investigating systematic dissociations between *experienced* utility, that is, the hedonic experience an option actually brings, and *decision* utility, the utility implied by the decision. Misprediction of experienced utility is common, as research on *affective forecasting* has repeatedly shown (Wilson & Gilbert, 2005). One of the most consistent findings is that people believe that future events (ranging from dramatic life changes to everyday occurrences, such as the outcomes of sporting events) will have a greater and longer lasting impact than they actually do (Gilbert, Pinel, Wilson, Blumberg, & Wheatley, 1998; Schkade & Kahneman, 1998; Wilson, Wheatley, Meyers, Gilbert, & Axsom, 2000). People also mispredict which experiences will make them happier, thus pursuing experiences that are ultimately less satisfying than they could be (Carlsmith, Wilson, & Gilbert, 2008; Wilson, Centerbar, Kermer, & Gilbert, 2005). A productive avenue for future research would be to investigate these affective forecasting errors with an eye toward helping people choose the experiences that will maximize their overall well-being.

Among other current and future directions, decision-making research is increasingly being enriched by a fuller consideration of the psychological processes that underlie judgment and choice phenomena. Researchers are using insights about attention, perception, memory, goal pursuit, and learning (to name just a few) to help them better understand how judgments and decisions are formed and why they sometimes prove inconsistent (Weber & Johnson, 2009). Along related lines, an active direction for decision-making research lies in the area of neuroeconomics, which gains insights into judgments and decisions by exploring their neural underpinnings (Loewenstein, Rick, & Cohen, 2008; see Camerer & Smith, Chapter 18). Not all current work on decision making is focusing increasingly “inward” on

the individual, however. Another active line of work is also exploring how established insights about decision making can be used to shape public policy decisions, so that behaviorally informed policies, which take into account decision makers’ natural tendencies and limitations (Barr, Mullainathan, & Shafir, 2008; Shafir, in press; Thaler & Sunstein, 2008), can be implemented.

A full description of human decision making needs to incorporate the issues discussed here as well as other tendencies not reviewed in this chapter, including various other judgment biases (see Griffin, Chapter 17), as well as people’s sensitivity to fairness (Kahneman, Knetsch, & Thaler, 1986a, 1986b; Rabin, 1993), aspiration levels (Lopes & Oden, 1999), and sunk costs (Arkes & Blumer, 1985; Gourville & Soman, 1998), among others. A successful descriptive model must allow for violations of normative criteria such as procedure and description invariance, dominance, regularity, and, occasionally, transitivity. It must also allow for the eventual incorporation of additional psychological processes that affect choice, as the refinement of descriptive theories is an evolving process. In any event, the product that emerges is, and will continue to be, quite distant from the elegant normative treatment. At the same time, acknowledged departures from the normative theory need not weaken that theory’s normative force. After all, normative theories are themselves empirical projects, capturing what, upon careful reflection, people consider ideal. As we improve our understanding of how decisions are made, we may be able to formulate prescriptive procedures to guide decision makers, in light of their limitations, to better capture their normative wishes.

Of course, there are instances where people have very clear preferences which no amount of subtle manipulation will alter (cf. Payne et al., 1992). At other times, we appear to be at the mercy of factors that we would prefer to consider inconsequential. This conclusion, well accepted within psychology, is becoming increasingly influential not only in decision research, but in public policy and the social sciences more generally. Prominent researchers in law, politics, medicine, sociology, and economics are exhorting their colleagues to pay attention to findings of the sort reviewed here in an attempt to formulate new ways of thinking about and predicting behavioral phenomena. Given the academic, personal, and applied import of decision making, these should be important developments for our understanding of why people act and decide as they do.

References

- Ainslie, G. (2001). *Breakdown of will*. New York: Cambridge University Press.
- Alter, A. L., & Oppenheimer, D. M. (2006). Predicting short-term stock fluctuations by using processing fluency. *Proceedings of the National Academy of Sciences*, 103, 9369–9372.
- Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, 13, 219–235.
- American Lung Association. (2011). Benefits of quitting. Retrieved August 2011, from <http://www.lungusa.org/stop-smoking/how-to-quit/why-quit/benefits-of-quitting/>
- Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin*, 110, 486–498.
- Arkes, H. R., & Blumer, C. (1985). The psychology of sunk cost. *Organizational Behavior and Human Decision Processes*, 35, 124–140.
- Arrow, K. J. (1951). Alternative approaches to the theory of choice in risk-taking situations. *Econometrica*, 19, 404–437.
- Arrow, K. J. (1988). Behavior under uncertainty and its implications for policy. In D. E. Bell, H. Raiffa, & A. Tversky (Eds.), *Decision making: Descriptive, normative, and prescriptive interactions* (pp. 497–507). Cambridge, England: Cambridge University Press.
- Barberis, N., Huang, M., & Thaler, R. H. (2006). Individual preferences, monetary gambles, and stock market participation: A case for narrow framing. *American Economic Review*, 96, 1069–1090.
- Barr, M. S., Mullainathan, S., & Shafir, E. (2008). *Behaviorally informed financial services regulation*. New America Foundation, Washington, DC. White Paper. (http://www.newamerica.net/files/naf_behavioral_v5.pdf)
- Bastardi, A., & Shafir, E. (1998). On the pursuit and misuse of useless information. *Journal of Personality and Social Psychology*, 75, 19–32.
- Bazerman, M. H., & Moore, D. A. (2008). *Judgment in managerial decision making* (7th ed.). Hoboken, NJ: Wiley.
- Bazerman, M. H., Schroth, H. A., Shah, P. P., Diekmann, K. A., & Tenbrunsel, A. E. (1994). The inconsistent role of comparison others and procedural justice in reactions to hypothetical job descriptions: Implications for job acceptance decisions. *Organizational Behavior and Human Decision Processes*, 60, 326–352.
- Benartzi, S., & Thaler, R. (1999). Risk aversion or myopia? Changes in repeated gambles and retirement investments. *Management Science*, 45, 364–381.
- Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica*, 22, 23–36. (Original work published in 1738).
- Berns, G. S., Laibson, D., & Loewenstein, G. (2007). Intertemporal choice – toward an integrative framework. *Trends in Cognitive Science*, 11, 482–488.
- Bettman, J. R., & Sujan, M. (1987). Effects of framing on evaluation of comparable and non-comparable alternatives by expert and novice consumers. *Journal of Consumer Research*, 14, 141–154.
- Bilgin, B., & LeBoeuf, R. A. (2010). Looming losses in future time perception. *Journal of Marketing Research*, 47, 520–530.
- Brenner, L., Rottenstreich, Y., Sood, S., & Bilgin, B. (2007). On the psychology of loss aversion: Possession, valence, and reversals of the endowment effect. *Journal of Consumer Research*, 34, 369–376.
- Buehler, R., Griffin, D., & Ross, M. (1994). Exploring the “planning fallacy”: Why people underestimate their task completion times. *Journal of Personality and Social Psychology*, 67, 366–381.
- Camerer, C. (1992). Recent tests of generalizations of expected utility theory. In W. Edwards (Ed.), *Utility theories: Measurement and applications* (pp. 207–251). Dordrecht, Netherlands: Kluwer.
- Camerer, C., Babcock, L., Loewenstein, G., & Thaler, R. (1997). Labor supply of New York City cabdrivers: One day at a time. *The Quarterly Journal of Economics*, 112, 407–441.
- Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, 19, 7–42.
- Camerer, C. F., Loewenstein, G., & Rabin, M. (Eds.). (2004). *Advances in behavioral economics*. Princeton, NJ: Princeton University Press.
- Carlsmith, K. M., Wilson, T. D., & Gilbert, D. T. (2008). The paradoxical consequences of revenge. *Journal of Personality and Social Psychology*, 95, 1316–1324.
- Chartrand, T. L., Huber, J., Shiv, B., & Tanner, R. J. (2008). Nonconscious goals and consumer choice. *Journal of Consumer Research*, 35, 189–201.
- Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82, 407–428.
- Connelly, T., Arkes, H. R., & Hammond, K. R. (Eds.). (2000). *Judgment and decision making: An interdisciplinary reader* (2nd ed.). Cambridge, England: Cambridge University Press.
- Cryder, C. E., Lerner, J. S., Gross, J. J., & Dahl, R. E. (2008). Misery is not miserly: Sad and self-focused individuals spend more. *Psychological Science*, 19, 525–530.
- Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34, 571–582.
- Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243, 1668–1674.
- Dawes, R. M., & Hastie, R. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision making*. Thousand Oaks, CA: Sage Publications.
- Downs, J. S., & Shafir, E. (1999). Why some are perceived as more confident and more insecure, more reckless and more cautious, more trusting and more suspicious, than others: Enriched and impoverished options in social judgment. *Psychonomic Bulletin and Review*, 6, 598–610.
- Elster, J., & Loewenstein, G. (Eds.). (1992). *Choice over time*. New York: Russell Sage Foundation.
- Fagley, N. S., & Miller, P. M. (1987). The effects of decision framing on choice of risky vs. certain options. *Organizational Behavior and Human Decision Processes*, 39, 264–277.
- Frederick, S., Loewenstein, G., & Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, 40, 351–401.
- Fujita, K., Trope, Y., Liberman, N., & Levin-Sagi, M. (2006). Construal levels and self-control. *Journal of Personality and Social Psychology*, 90, 351–367.
- Gilbert, D. T., Pinel, E. C., Wilson, T. D., Blumberg, S. J., & Wheatley, T. P. (1998). Immune neglect: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology*, 75, 617–638.

- Goldstein, W. M., & Hogarth, R. M. (Eds.). (1997). *Research on judgment and decision making: Currents, connections, and controversies*. New York: Cambridge University Press.
- Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, 38, 129–166.
- Gourville, J. T., & Soman, D. (1998). Payment depreciation: The behavioral effects of temporally separating payments from consumption. *Journal of Consumer Research*, 25, 160–174.
- Grether, D., & Plott, C. (1979). Economic theory of choice and the preference reversal phenomenon. *American Economic Review*, 69, 623–638.
- Harinck, F., Van Dijk, E., Van Beest, I., & Mersmann, P. (2007). When gains loom larger than losses: Reversed loss aversion for small amounts of money. *Psychological Science*, 18, 1099–1105.
- Heath, C., & Soll, J. B. (1996). Mental budgeting and consumer decisions. *Journal of Consumer Research*, 23, 40–52.
- Heath, C., & Tversky, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty*, 4, 5–28.
- Hershey, J. C., & Schoemaker, P. J. H. (1980). Risk taking and problem context in the domain of losses: An expected utility analysis. *The Journal of Risk and Insurance*, 47, 111–132.
- Hsee, C. K. (1996). The evaluability hypothesis: An explanation of preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes*, 67, 247–257.
- Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin*, 5, 576–590.
- Hsee, C. K., & Zhang, J. (2004). Distinction bias: Misprediction and mischoice due to joint evaluation. *Journal of Personality and Social Psychology*, 86, 680–695.
- Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research*, 9, 90–98.
- Isen, A. M., Nygren, T. E., & Ashby, F. G. (1988). Influence of positive affect on the subjective utility of gains and losses: It is just not worth the risk. *Journal of Personality and Social Psychology*, 55, 710–717.
- Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, 79, 995–1006.
- Johnson, E. J., & Goldstein, D. (2003). Do defaults save lives? *Science*, 302, 1338–1339.
- Johnson, E. J., Hershey, J., Meszaros, J., & Kunreuther, H. (1993). Framing, probability distortions, and insurance decisions. *Journal of Risk and Uncertainty*, 7, 35–51.
- Johnson, E. J., & Tversky, A. (1983). Affect, generalization, and the perception of risk. *Journal of Personality and Social Psychology*, 45, 20–31.
- Kahneman, D. (1992). Reference points, anchors, and mixed feelings. *Organizational Behavior and Human Decision Processes*, 51, 296–312.
- Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, 58, 697–720.
- Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 267–293). Cambridge, England: Cambridge University Press.
- Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986a). Fairness and the assumptions of economics. *Journal of Business*, 59, s285–s300.
- Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986b). Fairness as a constraint on profit seeking: Entitlements in the market. *American Economic Review*, 76, 728–741.
- Kahneman, D., Knetsch, J. L., & Thaler, R. (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economics*, 98, 1325–1348.
- Kahneman, D., & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science*, 39, 17–31.
- Kahneman, D., & Ritov, I. (1994). Determinants of stated willingness to pay for public goods: A study in the headline method. *Journal of Risk and Uncertainty*, 9, 5–38.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47, 263–291.
- Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, 39, 341–350.
- Kahneman, D., & Tversky, A. (1995). Conflict resolution: A cognitive perspective. In K. J. Arrow, R. H. Mnookin, L. Ross, A. Tversky, & R. B. Wilson (Eds.), *Barriers to conflict resolution* (pp. 45–60). New York: W.W. Norton and Company.
- Kahneman, D., & Tversky, A. (Eds.). (2000). *Choices, values, and frames*. Cambridge, England: Cambridge University Press.
- Keeney, R. L., & Raiffa, H. (1976). *Decisions with multiple objectives: Preferences and value tradeoffs*. Cambridge, England: Cambridge University Press.
- Keren, G. K. (2011). *Perspectives on framing*. London: Psychology Press.
- Knetsch, J. L. (1989). The endowment effect and evidence of nonreversible indifference curves. *American Economic Review*, 79, 1277–1284.
- Koehler, D. J., & Harvey, N. (2004). *Blackwell handbook of judgment and decision making*. Oxford, England: Blackwell.
- Kühberger, A. (1995). The framing of decisions: A new look at old problems. *Organizational Behavior and Human Decision Processes*, 62, 230–240.
- Kunreuther, H., & Pauly, M. (2005). Insurance decision-making and market behavior. *Foundations and Trends in Microeconomics*, 1, (monograph).
- LeBoeuf, R. A. (2006). Discount rates for time versus dates: The sensitivity of discounting to time-interval description. *Journal of Marketing Research*, 43, 59–72.
- LeBoeuf, R. A., & Shafir, E. (2003). Deep thoughts and shallow frames: On the susceptibility to framing effects. *Journal of Behavioral Decision Making*, 16, 77–92.
- LeBoeuf, R. A., Shafir, E., & Bayuk, J. B. (2010). The conflicting choices of alternating selves. *Organizational Behavior and Human Decision Processes*, 111, 48–61.
- Lerner, J. S., & Keltner, D. (2001). Fear, anger, and risk. *Journal of Personality and Social Psychology*, 81, 146–159.
- Lerner, J. S., Small, D. A., & Loewenstein, G. (2004). Heart strings and purse strings: Carryover effects of emotions on economic decisions. *Psychological Science*, 15, 337–341.
- Lerner, J. S., & Tiedens, L. Z. (2006). Portrait of the angry decision maker: How appraisal tendencies shape anger's influence on cognition. *Journal of Behavioral Decision Making*, 19, 115–137.
- Levin, I. P. (1987). Associative effects of information framing. *Bulletin of the Psychonomic Society*, 25, 85–86.
- Levin, I. P., & Chapman, D. P. (1990). Risk taking, frame of reference, and characterization of victim groups in AIDS

- treatment decisions. *Journal of Experimental Social Psychology*, 26, 421–434.
- Levin, I. P., Gaeth, G. J., Schreiber, J., & Lauriola, M. (2002). A new look at framing effects: Distribution of effect sizes, individual differences, and independence of types of effects. *Organizational Behavior and Human Decision Processes*, 88, 411–429.
- Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes*, 76, 149–188.
- Levin, I. P., Schnittjer, S. K., & Thee, S. L. (1988). Information framing effects in social and personal decisions. *Journal of Experimental Social Psychology*, 24, 520–529.
- Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology*, 89, 46–55.
- Lichtenstein, S., & Slovic, P. (1973). Response-induced reversals of preferences in gambling: An extended replication in Las Vegas. *Journal of Experimental Psychology*, 101, 16–20.
- Lichtenstein, S., & Slovic, P. (2006). *The construction of preference*. Cambridge, England: Cambridge University Press.
- Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes*, 65, 272–292.
- Loewenstein, G., & Prelec, D. (1992). Anomalies in intertemporal choice: Evidence and an interpretation. *The Quarterly Journal of Economics*, 107, 573–597.
- Loewenstein, G., Rick, S., & Cohen, J. D. (2008). Neuroeconomics. *Annual Review of Psychology*, 59, 647–672.
- Loewenstein, G., & Thaler, R. H. (1989). Intertemporal choice. *Journal of Economic Perspectives*, 3, 181–193.
- Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, 127, 267–286.
- Lopes, L. L., & Oden, G. C. (1999). The role of aspiration level in risky choice: A comparison of cumulative prospect theory and SP/A theory. *Journal of Mathematical Psychology*, 43, 286–313.
- Madrian, B. C., & Shea, D. F. (2001). The power of suggestion: Inertia in 401(k) participation and savings behavior. *The Quarterly Journal of Economics*, 116, 1146–1187.
- Mandel, N., & Johnson, E. J. (2002). When web pages influence choice: Effects of visual primes on experts and novices. *Journal of Consumer Research*, 29, 235–245.
- McGraw, A. P., Shafir, E., & Todorov, A. (2010). Valuing money and things: Why a \$20 item can be worth more and less than \$20. *Management Science*, 56, 816–830.
- McNeil, B. J., Pauker, S. G., Sox, H. C., & Tversky, A. (1982). On the elicitation of preferences for alternative therapies. *New England Journal of Medicine*, 306, 1259–1262.
- Morris, M. W., Carranza, E., & Fox, C. R. (2008). Mistaken identity: Activating conservative political identities induces “conservative” financial decisions. *Psychological Science*, 19, 1154–1160.
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231–259.
- Novemsky, N., Dhar, R., Schwarz, N., & Simonson, I. (2007). Preference fluency in choice. *Journal of Marketing Research*, 44, 347–356.
- Novemsky, N., & Kahneman, D. (2005). The boundaries of loss aversion. *Journal of Marketing Research*, 42, 119–128.
- Payne, J. W., Bettman, J. R., & Johnson, E. J. (1992). Behavioral decision research: A constructive processing perspective. *Annual Review of Psychology*, 43, 87–131.
- Peetz, J., & Buehler, R. (2009). Is there a budget fallacy? The role of savings goals in the prediction of personal spending. *Personality and Social Psychology Bulletin*, 35, 1579–1591.
- Prelec, D. (2000). Compound invariant weighting functions in prospect theory. In D. Kahneman & A. Tversky (Eds.), *Choices, values, and frames* (pp. 67–92). New York: Cambridge University Press.
- Quattrone, G. A., & Tversky, A. (1988). Contrasting rational and psychological analyses of political choice. *American Political Science Review*, 82, 719–736.
- Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, 83, 1281–1302.
- Read, D., Frederick, S., Orsel, B., & Rahman, J. (2005). Four score and seven years from now: The “date/delay” effect in temporal discounting. *Management Science*, 51, 1326–1335.
- Redelmeier, D. A., & Shafir, E. (1995). Medical decision making in situations that offer multiple alternatives. *Journal of the American Medical Association*, 273, 302–305.
- Redelmeier, D., Shafir, E., & Aujla, P. (2001). The beguiling pursuit of more information. *Medical Decision Making*, 21, 376–381.
- Redelmeier, D. A., & Tversky, A. (1990). Discrepancy between medical decisions for individual patients and for groups. *New England Journal of Medicine*, 322, 1162–1164.
- Redelmeier, D. A., & Tversky, A. (1992). On the framing of multiple prospects. *Psychological Science*, 3, 191–193.
- Roszkowski, M. J., & Snelbecker, G. E. (1990). Effects of framing on measures of risk tolerance: Financial planners are not immune. *Journal of Behavioral Economics*, 19, 237–246.
- Rottenstreich, Y., & Hsee, C. K. (2001). Money, kisses, and electric shocks: On the affective psychology of risk. *Psychological Science*, 12, 185–190.
- Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1, 7–59.
- Savage, L. J. (1954). *The foundations of statistics*. Wiley: New York.
- Sayette, M. A., Loewenstein, G., Griffin, K. M., & Black, J. J. (2008). Exploring the hot-to-cold empathy gap in smokers. *Psychological Science*, 19, 926–932.
- Schkade, D. A., & Johnson, E. J. (1989). Cognitive processes in preference reversals. *Organizational Behavior and Human Decision Processes*, 44, 203–231.
- Schkade, D. A., & Kahneman, D. (1998). Does living in California make people happy? A focusing illusion in judgments of life satisfaction. *Psychological Science*, 9, 340–346.
- Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory and Cognition*, 21, 546–556.
- Shafir, E. (1994). Uncertainty and the difficulty of thinking through disjunctions. *Cognition*, 50, 403–430.
- Shafir, E. (2002). Cognition, intuition, and policy guidelines. In R. Gowda & J. C. Fox (Eds.), *Judgments, decisions, and public policy* (pp. 71–88). New York: Cambridge University Press.
- Shafir, E. (in press). *The behavioral foundations of policy*. Princeton, NJ: Princeton University Press.
- Shafir, E., Diamond, P., & Tversky, A. (1997). Money illusion. *Quarterly Journal of Economics*, 112, 341–374.
- Shafir, E., & LeBoeuf, R. A. (2002). Rationality. *Annual Review of Psychology*, 53, 491–517.
- Shafir, E., & LeBoeuf, R. A. (2004). Context and conflict in multiattribute choice. In D. Koehler & N. Harvey (Eds.),

- Blackwell handbook of judgment and decision making* (pp. 341–359). Malden, MA: Blackwell.
- Shafir, E., Simonson, I., & Tversky, A. (1993). Reason-based choice. *Cognition*, 49, 11–36.
- Shafir, E., & Tversky, A. (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology*, 24, 449–474.
- Sieck, W., & Yates, J. F. (1997). Exposition effects on decision making: Choice and confidence in choice. *Organizational Behavior and Human Decision Processes*, 70, 207–219.
- Simmons, J. P., & Nelson, L. D. (2006). Intuitive confidence: Choosing between intuitive and nonintuitive alternatives. *Journal of Experimental Psychology: General*, 135, 409–428.
- Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, 69, 99–118.
- Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects. *Journal of Consumer Research*, 16, 158–174.
- Simonson, I., & Tversky, A. (1992). Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research*, 29, 289–295.
- Slovic, P. (1995). The construction of preference. *American Psychologist*, 50, 364–371.
- Slovic, P., Finucane, M., Peters, E., & MacGregor, D. G. (2002). The affect heuristic. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 397–420). New York: Cambridge University Press.
- Slovic, P., Fischhoff, B., & Lichtenstein, S. (1988). Response mode, framing, and information-processing effects in risk assessment. In D. E. Bell, H. Raiffa, & A. Tversky (Eds.), *Decision making: Descriptive, normative, and prescriptive interactions* (pp. 152–166). Cambridge, England: Cambridge University Press.
- Slovic, P., Griffin, D., & Tversky, A. (1990). Compatibility effects in judgment and choice. In R. M. Hogarth (Ed.), *Insights in decision making: A tribute to Hillel J. Einhorn* (pp. 5–27). Chicago, IL: University of Chicago Press.
- Smith, V. L. (1985). Experimental economics: Reply. *American Economic Review*, 75, 265–272.
- Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.
- Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, 94, 672–695.
- Sunstein, C. R., Kahneman, D., Schkade, D., & Ritov, I. (2002). Predictably incoherent judgments. *Stanford Law Review*, 54, 1153–1215.
- Thaler, R. H. (1981). Some empirical evidence on dynamic inconsistency. *Economic Letters*, 8, 201–207.
- Thaler, R. H. (1991). The psychology of choice and the assumptions of economics. In R. H. Thaler (Ed.), *Quasi-rational economics* (pp. 137–166). New York: Russell Sage Foundation.
- Thaler, R. H. (1993). *Advances in behavioral finance*. New York: Russell Sage Foundation.
- Thaler, R. H. (1999). Mental accounting matters. *Journal of Behavioral Decision Making*, 12, 183–206.
- Thaler, R. H., & Sunstein, C. R. (2008). *Nudge*. New Haven, CT: Yale University Press.
- Trope, Y., & Liberman, N. (2003). Temporal construal. *Psychological Review*, 110, 403–421.
- Trope, Y., Liberman, N., & Wakslak, C. (2007). Construal levels and psychological distance: Effects on representation, prediction, evaluation, and behavior. *Journal of Consumer Psychology*, 17, 83–95.
- Turner, J. C. (1985). Social categorization and the self-concept: A social cognitive theory of group behavior. In E. J. Lawler (Ed.), *Advances in group processes* (Vol. 2, pp. 77–121). Greenwich, CT: JAI Press.
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and psychology of choice. *Science*, 211, 453–458.
- Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of Business*, 59, s251–s278.
- Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference dependent model. *Quarterly Journal of Economics*, 106, 1039–1061.
- Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5, 297–323.
- Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review*, 95, 371–384.
- Tversky, A., & Shafir, E. (1992a). Choice under conflict: The dynamics of deferred decision. *Psychological Science*, 3, 358–361.
- Tversky, A., & Shafir, E. (1992b). The disjunction effect in choice under uncertainty. *Psychological Science*, 3, 305–309.
- Tversky, A., & Simonson, I. (1993). Context-dependent preferences. *Management Science*, 39, 1178–1189.
- Tversky, A., Slovic, P., & Kahneman, D. (1990). The causes of preference reversal. *American Economic Review*, 80, 204–217.
- Van Boven, L., Dunning, D., & Loewenstein, G. (2000). Egocentric empathy gaps between owners and buyers: Misperceptions of the endowment effect. *Journal of Personality and Social Psychology*, 79, 66–76.
- Verplanken, B., & Holland, R. W. (2002). Motivated decision making: Effects of activation and self-centrality of values on choices and behavior. *Journal of Personality and Social Psychology*, 82, 434–447.
- von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.
- Weber, E. U., & Johnson, E. J. (2009). Mindful judgment and decision making. *Annual Review of Psychology*, 60, 53–85.
- Wedell, D. H. (1997). Another look at reasons for choosing and rejecting. *Memory and Cognition*, 25, 873–887.
- Wegner, D. M., & Bargh, J. A. (1998). Control and automaticity in social life. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., Vol. 1, pp. 446–496). Boston, MA: McGraw-Hill.
- Wheeler, S. C., & Berger, J. (2007). When the same prime leads to different effects. *Journal of Consumer Research*, 34, 357–368.
- Wilson, T. D., Centerbar, D. B., Kermel, D. A., & Gilbert, D. T. (2005). The pleasures of uncertainty: Prolonging positive moods in ways people do not anticipate. *Journal of Personality and Social Psychology*, 88, 5–21.
- Wilson, T. D., Dunn, D. S., Kraft, D., & Lisle, D. J. (1989). Introspection, attitude change, and attitude consistency: The disruptive effects of explaining why we feel the way we do. In L. Berkowitz (Ed.), *Advances in experimental and social psychology* (pp. 123–205). San Diego, CA: Academic Press.
- Wilson, T. D., & Gilbert, D. T. (2005). Affective forecasting: Knowing what to want. *Current Directions in Psychological Science*, 14, 131–134.

- Wilson, T. D., Lisle, D. J., Schooler, J. W., Hodges, S. D., Klaaren, K. J., & LaFleur, S. J. (1993). Introspecting about reasons can reduce post-choice satisfaction. *Personality and Social Psychology Bulletin, 19*, 331–339.
- Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology, 60*, 181–192.
- Wilson, T. D., Wheatley, T., Meyers, J. M., Gilbert, D. T., & Axsom, D. (2000). Focalism: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology, 78*, 821–836.
- Zauberman, G., Kim, B. K., Malkoc, S. A., & Bettman, J. R. (2009). Discounting time and time discounting. *Journal of Marketing Research, 46*, 543–556.

Judgmental Heuristics: A Historical Overview

Dale W. Griffin, Richard Gonzalez, Derek J. Koehler, and Thomas Gilovich

Abstract

The Heuristics and Biases approach to judgment under uncertainty began 40 years ago with the publication of a study of the statistical foibles on the part of research psychologists and statisticians (Tversky & Kahneman, 1971). Since then, this research program has substantially influenced the working assumptions of psychologists and economists about the role of normative models of probability judgment and decision making, while providing a new language of judgmental heuristics. We provide a historical overview of the Heuristics and Biases research program that describes its intellectual antecedents and the special role of the rational actor model in shaping its methods, and we review the program's evolution over the course of three waves of research and theory development.

Key Words: judgment, heuristics, biases, rationality, subjective probability, representativeness, availability, anchoring, adjustment

“Predictions are difficult to make, especially about the future.” This statement, attributed by different sources to a United Nations official, Niels Bohr, and Yogi Berra, may be taken as self-excusing, self-mocking, or simply confused. Although it is difficult to consider all relevant factors when evaluating the probability of a sports team winning, a stock increasing in value, or a relationship leading to marriage, when we consider such matters—*even briefly*—a feeling of certainty or uncertainty seems to “pop out.” For example, when a respected British politician was asked whether Kosovo peace talks would lead to a settlement, he stated—with confidence and after only a brief pause—that “the balance of probabilities are 40–60 against.” How did he do that?

According to the “Heuristics and Biases” (H&B) approach to human judgment, people typically use cognitive shortcuts that make assessments of likelihood quick and easy but prone to systematic error.

Such shortcuts occur not only in predictions but in retrospective judgments of probability as well, and they can be recognized through signature “biases,” as we describe later. Consider a recent article in a major national newspaper. The article, titled the “20 million to 1 family,” described how a couple had “broken all records by having eight children born in symmetrical girl-boy, girl-boy, girl-boy, girl-boy order.” The explanation of this strange rendering of the odds based on judgmental heuristics is that people incorrectly (but easily and effortlessly) judge the target sequence of births to be extremely unlikely because the symmetrical pattern of births does not match and is *not representative* of a random series. Formal probability theory, in contrast, prescribes that any sequence of four boys and four girls is as likely as any other.

Based partly on their experience teaching statistics and on their observations of judgments and predictions in applied settings, Daniel Kahneman and

Amos Tversky (Kahneman & Tversky, 1972, 1973; Tversky & Kahneman, 1971, 1974) proposed that intuitive judgments under uncertainty are typically controlled by judgmental “heuristics” rather than by the formal laws of probability. Kahneman and Tversky were not the first to suggest that classical “rational” models of statistical reasoning fail to describe actual human reasoning in many settings, but their program of research was both more radical and more influential than most others. Their challenge to rational models influenced theory and research not only in cognitive psychology but also in social psychology, economics, philosophy, political science, medical decision making, and legal studies.

We first discuss the meaning of “rationality” that is most relevant to the heuristics and biases program, review the negative and positive messages of the original program, explore the chief criticisms of that program, and finally present extensions to the original heuristics and biases research. Our presentation is historical in focus and organization, and readers seeking a more focused treatment of recent reconceptualizations are advised to consult Kahneman and Frederick’s (2005) chapter in the previous edition of this Handbook.

The Rational Model

The classical model of rational choice (see Chater & Oaksford, Chapter 2) is central to the discipline of economics, and at its heart is the guiding principle of maximizing subjective expected utility (SEU). According to this model, which provides a behavioral definition and measure of rationality, the “rational actor” assesses the attractiveness of a given option by evaluating the probability of each possible resultant outcome and combining that subjective probability with the subjective utility or personal value of each outcome. The rational economic actor then chooses the best option on the basis of the optimal probability-weighted utility. Economic theories that guide public policy in areas as diverse and important as taxation, environmental safety, stock market and banking regulation, and Social Security rely on the central assumption that individuals and organizations are rational in this sense. Underpinning this model is a series of axioms, or simple rules of logic, that are defined to be both intuitively and formally compelling in their abstract form. This axiomatic foundation provides a series of sharp tests that clearly assess the degree to which observed judgments fit this specific (and widely applied) rational model. The behavioral findings

of Kahneman and Tversky (and many colleagues) question the fundamental assumptions of this normative model by showing how these axiomatic tests fail under well-specified conditions.

There are many events for which it is easy to calculate the “correct” probability (e.g., the chance of drawing a given hand of cards). But in other cases, such as the prediction of peace in our time, the appropriateness of the probability judgment can only be tested by examining its *coherence* relative to other judgments (e.g., the probability of a subcategory must be less than or equal to its superordinate category), and by examining its *calibration* when aggregated together with several other judgments equated on probability (i.e., events predicted with .70 probability must occur 70% of the time). Note that coherence can be satisfied with respect to purely internal criteria, whereas calibration is specifically defined with respect to external criteria: how many things actually happened in the world. Violations of rationality in this model, then, do not imply anything about the relative importance of “hot” emotional versus “cool” cognitive factors; by this definition, rationality requires only that people follow the rules of subjective probability and evaluate their own preferences consistently.

The most widely used benchmark of the coherence of probability assessment is Bayes’ rule, which has been described as the “master rule” of categorical inference or categorical prediction (see Fischhoff & Beyth-Marom, 1983, for an early psychologically oriented discussion of Bayesian hypothesis testing; also Griffiths, Tenenbaum, & Kemp, Chapter 3). Bayes’ rule defines how to use probability theory to update the probability of a hypothesis given some data. For example, when inferring the probability that a patient has heart disease (H1) on the basis of a positive diagnostic test (D), a rational physician would (implicitly or explicitly) calculate the following quantity, where H2 refers to the probability that the patient does not have heart disease.

$$\frac{P(H1|D)}{P(H2|D)} = \frac{P(D|H1)}{P(D|H2)} * \frac{P(H1)}{P(H2)}$$

The first quantity on the right-hand side is the likelihood ratio, which expresses the *relative likelihood* that a patient known to have heart disease would yield the test result D (for data) compared to a patient known not to have heart disease. The likelihood ratio thus reflects the *diagnosticity* of the given evidence D. In general, diagnosticity increases

with increasing separability of the two competing hypotheses, increasing quality of the diagnostic data, and increasing sample size of the diagnostic data. For example, a given blood pressure reading would be more diagnostic in distinguishing between heart disease and a healthy heart than between heart disease and another vascular disease; it would be more diagnostic if it were taken by an experienced physician than by a beginning medical student; and it would be more diagnostic if it were based on the average of many readings than based on a single reading. The second quantity on the right-hand side is the prior odds ratio, which reflects the relative prevalence of the two outcomes in the relevant population, that is, the relative probability of encountering a given member of each class (in the frequentist approach to probability, the chance of encountering a given member in one of many random draws).

The strength of inference that can be drawn from a given body of evidence depends on the relative balance of the likelihood ratio and the prior odds ratio. If, for example, the diagnostic test has good validity such that the likelihood ratio is 9:1 in favor of heart disease given a positive test result, then a prior odds ratio of 1:9 against heart disease leaves the rational physician with posterior odds of 1:1, or a .5 probability that the patient has heart disease. If, on the other hand, prior odds of 1:9 against are matched by a likelihood ratio of 1:9 against, then the posterior odds are 1:81 against, or a little over a .01 probability that the patient has heart disease.

The use of Bayes' rule to describe "ideal" probabilistic judgment in frequentistic settings, with repeated, exchangeable events such as drawing balls from an urn, is entirely uncontroversial. However, when Bayes' rule is used to prescribe the updating of subjective probabilities about a unique event, some controversy entails (e.g., Savage, 1954). In particular, some statisticians argue that probability theory can only be applied to the frequentist case. However, as many applied researchers (including Keynes, 1921) have argued, if probabilistic statements about unique, real-world events are excluded from the domain of probability theory, nothing interesting is left. Wars, depressions, mergers, marriages, deaths and divorces may happen with some regularity, but each is experienced as a unique event. Are probability judgments about such events without guidelines or standards? For now, it is enough that the classical economic model of rationality—using the principle of coherence—requires subjective probability judgments to follow Bayes' rule.

Historical Antecedents of the Heuristics and Biases Program

In the 1950s, inspired by the use of expert judgment in engineering systems developed during World War II, by the cognitive revolution that required human judgment to be modeled in terms of computer systems, and by the increasing contact between experimental psychology and economic decision-making models, a number of research programs examined the issues of coherence and calibration in human probabilistic judgment. Herbert Simon (1957), early in his Nobel Prize-winning research on economic models, argued that "full" rationality was an unrealistic assumption because of processing limitations in living systems (and, incidentally, in virtually all computers currently available). He proposed a limited form of rationality, termed "bounded rationality," that accepted the limited search and computational ability of human brains but nonetheless assumed that after a truncated search and after considering a limited subset of alternatives, people did act and reason rationally, at least in terms of achieving their goals.

It is worthwhile digging deeper into the Simonian critique, as Simon borrowed the use of the term "heuristic" from computer science and artificial intelligence to describe simplified yet highly efficient human reasoning principles, setting the stage for all future work on heuristics of judgment and decision making. Simon was trained in the field of public administration, and he was originally interested in modeling how bureaucracies worked (a goal more focused on "description" than on "prescription"). The phenomena that Simon and his colleagues observed could be described as "muddling through"—large organizations seemed to operate on simple rules of thumb in an environment in which no one person or department knows everything but somehow everyone knows just enough to produce an adequate outcome. Later, he turned his attention to the psychology of problem solving, with a particular interest in expert judgment. Simon did not build his theories of bounded rationality on specific psychological principles or processes: He explicitly noted that psychological theories of choice processes were not yet sufficiently developed to inform economics. Instead he used general psychological principles to outline some broad, realistic constraints on rational models as models of actual decision making. These general psychological principles reflected the zeitgeist of cognitive psychology at the time, which focused on the limits of memory and attention.

Simon's realistic constraints set the stage for the study of judgmental heuristics and the field of judgment and decision making more generally. In his theories of bounded rationality, he asserted that people simplify the choice process by searching for a satisfactory rather than an optimal outcome. *Satisficing*, he argued, generally consists of three elements: a strategy that examines local or easy options before looking further afield, a stopping rule that specifies an *aspiration level* that must be met and hence how far afield the search should continue, and a simplified assessment of future value that provides a rather vague clue as to the actual value of the choice. There is another less well-known side to Simon's critique: He also maintained that such simplified methods of choice can do surprisingly well relative to optimizing methods and that "bounded rationality" could still be evolutionarily successful.

Simon offered the field of economics (at least) two other familiar psychological insights that were to echo repeatedly in the development of behavioral models of judgment and decision making. First, the human mind (as well as the aggregate mind of the organization) can only hold on to two or three alternatives at one time. Second, attention is a precious and costly commodity, a fact that must be considered in any description of how judgment and choice processes actually operate. Thus, in the vocabulary later introduced by Kahneman and Tversky, Simon had both a negative agenda (explaining how ideal, rational models were unrealistic and descriptively invalid) and a positive agenda (providing guidelines as to how humans—and animals—might actually make highly sensible, if simplified, choices).

Research by Ward Edwards (reviewed in Edwards, 1968) was designed to test rationality assumptions more directly. Using bags full of colored poker chips to explore how people revised, or "updated," their probabilities in the face of new evidence, Edwards concluded that people are not always well calibrated (that is, their probability judgments are not accurate but biased) but are generally coherent in their judgments. In particular, he and his colleagues concluded that *in general* people do reason in accordance with the rules of probability (as summarized by Bayes' rule); however, they give new evidence too little weight and thus are "conservative." It is important to our later arguments to note that conservatism was only the most *common* finding in this research program. Systematic exceptions were found when participants were given new evidence of low probative weight; in this case, judgments were

typically "radical," giving too much weight to the new evidence.

The work by Simon and by Edwards and colleagues is generally seen as the main predecessor of the H&B approach. However, there were several other flourishing research programs on subjective probability in the 1950s and 1960s that cast further doubt on the rationality assumption. For example, Adams and Adams (1961) examined the calibration of subjects' probability judgments about their own knowledge and found consistent "overconfidence:" For most probability levels, the actual percentage of correct answers to general knowledge questions was too low to justify the judged probability of being correct. Researchers using the Signal Detection model (which also has a Bayesian foundation) to study human perceptual judgments (e.g., Pollack & Decker, 1958) found that the correspondence between the rated probability of a "signal" being present and its actual probability depended on the difficulty of the recognition problem. When the stimulus was relatively difficult to recognize (e.g., because a tone was degraded with random noise or because a visual stimulus was very small), receivers' subjective probability judgments were too close to 1.0, that is, they were overconfident. When the stimulus was relatively easy to recognize, receivers' subjective probability judgments corresponded closely to the actual probability of receiving a signal and sometimes were even too low.

Throughout the 1950s, J. Cohen (e.g., Cohen & Hansel, 1956) studied intuitive conceptions of probability in children and adults, especially in terms of belief in "chance" and "luck" in gambling and risk-taking behavior. He concluded that intuitive conceptions of probability were qualitatively different than those described by the axioms of probability theory. Anomalies in conceptions of randomness noted by Cohen and others included two particularly robust phenomena: the gambler's fallacy and probability matching. The gambler's fallacy is the belief (implicit or explicit) that the "law of averages" requires that the probability of a given outcome of a chance device (e.g., Tails when tossing a coin) increases with a run of the alternate outcome (e.g., tossing Heads many times). Probability matching is the practice of predicting the more common event in proportion to the base-rate frequency of that event (e.g., if a roulette wheel was designed to end up "red" on 70% of spins, a probability-matching bettor would bet "red" on 70% of the trials, instead of betting "red" on *every* trial, which would maximize the expected number of wins).

About the same time, Paul Meehl was describing two fundamental challenges to the optimality of clinical judgment. First, he noted that clinical prediction was almost entirely based on characteristics of the case being judged with little or no concern for the relative prevalence or “base rates” of the possible outcomes (Meehl & Rosen, 1955). Second, he compiled a list of studies that compared the accuracy of clinical prediction with actuarial or formula-based prediction: Formulas did better (Meehl, 1954). Some time afterward, Oskamp (1965) demonstrated how trained clinical judges become increasingly miscalibrated (overconfident) as they gained more data about a case. Later, Mischel (1968) challenged the validity of clinical interviews to predict future behavior in very different situations. Most important for the present review, he pointed to the discrepancy between judges’ beliefs and the empirical evidence of poor predictive validity.

These diverse findings and perspectives set the stage for Kahneman and Tversky’s judgmental heuristics account of intuitive probability. The H&B program was not a comprehensive attempt to explain the anomalies that littered the field of human judgment, but naturally it was influenced by what came before. It was an attempt to describe some of the most notable elements of human judgment that Kahneman and Tversky observed in the classroom and in the real world. Simon and Edwards had brought the potential conflict between normative rational models and descriptive human models into sharp focus, but they had concluded that people were approximately or boundedly rational, within limits determined by their computational capacity. However, there was considerable evidence that the assumption of calibration was generally untenable, and some evidence from Cohen’s work that the axioms that required coherence were not consistent with intuitive judgments of probability. In this context, Kahneman and Tversky took a radical step: They proposed that the rules of probability, which define the rational “best guess” about outcomes, are not natural or intuitive methods of assessing degrees of belief or likelihood. Furthermore, they implied, simplifying the search set or restricting the number of computations was not enough to rescue the rationality assumptions. Instead, in many situations people naturally and spontaneously assess the likelihood of an outcome by processes that are qualitatively different from the rules of probability theory. In other words, “intuitive” judgment is not boundedly

rational, but not rational at all (at least in the classical “rational actor” sense).

Later critics have argued that the H&B program marked a sudden and arbitrary shift away from prior research on conservatism, which largely upheld the assumption of rationality (e.g., Gigerenzer & Murray, 1987; Lopes, 1991). This criticism is hard to support, for, as we explain later, the H&B model is consistent with conservatism as well as with the other anomalies listed earlier. The H&B program accounted for the previous findings and also predicted many specific laboratory-based anomalies presented and tested in Kahneman and Tversky’s early papers. We must emphasize that the laboratory-based demonstrations were never meant to be the phenomena to be explained—they were meant to illustrate and test the *processes* thought to underlie the real phenomena of interest and to illuminate specific tests that could sharply reject the behavioral applicability of the underlying axioms. The phenomena to be explained were judgments in the real world that seemed to be at odds with the dictates of probability theory.

Three Heuristics Explain Many Biases: The First Wave of Research

The Heuristics and Biases program began when Amos Tversky, a mathematical psychologist who had worked with Edwards and others on formal measurement models, described the current state of the Behavioral Decision Theory paradigm circa 1968 to Daniel Kahneman, his colleague in the Psychology Department at Hebrew University. Kahneman found the idea of tinkering with formal models such as SEU to make them fit the accumulating empirical evidence to be an unpromising approach to understanding the psychological processes involved in judgment and choice. Instead, he argued, based on his own research on visual attention and processing, the principles of cognition underlying judgment should follow the principles of perception (cf. Brunswik, 1956). Thus, instead of starting with formal models as the basis of descriptive accounts of judgment and decision making, Kahneman and Tversky started with principles of perception and psychophysics and extended them to the kind of processing necessary to evaluate probabilities and assess subjective values.

This approach immediately suggested a guiding paradigm for research on judgment and decision making: the study of visual illusions. The logic of studying perceptual illusions is that failures of a

system are often more diagnostic of the rules the system follows than are its successes. Consider, for example, the moon illusion: The full moon looks enormous as it sits on the horizon, but it appears more modestly sized when high in the sky. There is little to learn from the constancy of the perceived size of the moon along the long arc of the overhead sky, but its illusory magnification when it sits on the horizon provides insight about the way that the visual system uses contextual detail to compute perceived distance and hence perceived size. The visual illusion paradigm, like the cognitive illusion approach patterned on it, does not imply that judgments of size are typically wrong—in fact, it provides a map to those situations when intuitive perceptions are likely to be correct—but it highlights the processes by which perception or judgment is *constructed* from imperfect cues. We would not say that the visual system is “irrational” because it uses environmental cues in a heuristic way; rather, we can conclude that using environmental cues in a heuristic way gives rise—in well-specified, but not necessarily common circumstances—to systematic and diagnostic errors or biases.

Thus, the guiding logic of Kahneman and Tversky’s approach to the study of judgment was in practice the opposite of that championed by Simon, who had urged researchers to seek out and understand the environmental factors that maximized the success of simple processes. The cognitive illusion paradigm seeks out those environments or problem descriptions in which the judgment and choice processes people rely on lead to clear errors. The purpose was not to emphasize the predominance of bias over accuracy, but to find the clearest testing grounds for diagnosing the underlying simple processes or *judgmental heuristics* that people habitually employ. This is an important distinction that many subsequent critiques have failed to appreciate, and it is worth quoting Kahneman and Tversky’s original description of the logic of the H&B paradigm:

The subjective assessment of probability resembles the subjective assessment of physical quantities such as distance or size. These judgments are all based on data of limited validity, which are processed according to heuristic rules. For example, the apparent distance of an object is determined in part by its clarity. The more sharply the object is seen, the closer it appears to be. This rule has some validity, because in any given scene the more distant objects are seen less sharply than nearer objects. However, the

reliance on this rule leads to systematic errors in the estimation of distance. Specifically, distances are often over-estimated when visibility is poor because the contours of objects are blurred. On the other hand, distances are often underestimated when visibility is good because the objects are seen sharply. Thus the reliance on clarity as an indication of distance leads to common biases. Such biases are also found in intuitive judgments of probability. (Tversky & Kahneman, 1974, reprinted in Kahneman, Slovic, & Tversky, 1982, p. 3)

The heuristics that Kahneman and Tversky identified were also suggested by the principles of perceptual psychology, especially the organizing principles of Gestalt psychology (e.g., Koffka, 1935). Gestalt psychology emphasized how the perceptual system effortlessly and without awareness creates whole forms even when the information reaching the receptors is incomplete and indeterminate. According to the H&B approach, these underlying heuristics are not a simplified version of an ideal statistical analysis but something completely different. This constituted a key point of differentiation between the H&B account and others before it: “In his evaluation of evidence, man is apparently not a conservative Bayesian: he is not Bayesian at all” (Kahneman & Tversky, 1972, p. 450). Unfortunately, or so it seems to us, this statement was taken by some to imply that the H&B (hu)man was not simply un-Bayesian, but rather stupid.

In a later phase of their collaborative research, Kahneman and Tversky took the perceptual framework they had used to study probability judgment and used it to illuminate decision making under risk, leading to their most complete and formal model, Prospect Theory (Kahneman & Tversky, 1979), for which Kahneman received the Nobel Prize in Economics in 2002. In this model, fundamental perceptual principles such as comparison levels and adaptation (Helson, 1964), diminishing sensitivity, and the privileged status of pain served as the primitives of a model that once again used specific biases and errors as tools of diagnosis (see LeBoeuf & Shafir, Chapter 16).

It is illuminating to compare the evolutionary implications of Simon’s Bounded Rationality and the H&B approach. For Simon, the guiding evolutionary principle was computational realism (i.e., simplified approximation) that nonetheless was well adapted to fit the information environment. For Kahneman and Tversky, the guiding evolutionary

principle was that existing processes in perceptual analysis were coopted as tools for higher level cognitive processing. Although these tools might work well in many environments, they also lead to signature biases that are endemic to human intuition. In many cases, the biases that Kahneman and Tversky were signals of underlying heuristics were already well known. As noted earlier, Meehl and Rosen (1955) had warned clinicians of the danger of neglecting base rates in psychological diagnoses. In other cases, the biases were identified by informal observation, whether of psychologists who seemed to neglect power and underestimate sample sizes, army officers who neglected regression effects in determining the value of rewards versus punishment, or army selection personnel who maintained their belief in the efficacy of interviews despite statistical evidence to the contrary.

Negative and Positive Aspects of the Heuristics Program

From the first articles on heuristics and biases, Kahneman and Tversky noted that their program had two interrelated messages, one negative, about how intuitions *do not* work, and one positive, about how intuitions *do* work. In retrospect, it seems possible to identify two or three distinct stages of the program. In the first stage, the focus was on the surface structure of judgmental heuristics, and demonstrations were designed to show how case-specific information dominated intuitive judgment and led, at times, to the complete neglect of other normatively important information. The second stage (or as we describe it later, the “second wave”) attempted to describe the deep psychological structure of judgmental heuristics, and the accompanying demonstrations were more likely to show how the (often conflicting) multiple sources of information were weighted. Finally, the third stage organized a broader set of heuristic processes under the rubric of a dual-process model of reasoning and judgment.

In the first stage, which dates from the original collaboration in 1969 to the 1974 summary paper, Kahneman and Tversky focused primarily on defining three judgmental heuristics (representativeness, availability, and anchoring and adjustment) by means of analogies with the processes underlying perceptual illusions. In simple, between-subject scenario experiments, Kahneman and Tversky demonstrated that people neglect prior odds (“base rates”), sample size, evidence quality, and diagnosticity, and instead rely on their immediate evaluation of the

strength of the sample evidence to construct their subjective probability judgments. The experiments focused on everyday judgments and predictions about hospital births, school achievement, and professional membership, rather than abstract textbook probability questions about balls and urns, or dice and coins. Such a shift in context was neither irrelevant nor unplanned, as the authors noted that questions about chance devices were most likely to trigger the use of statistical rules rather than intuitive thinking. The authors acknowledged that almost any reasoning problem could be made “transparent” enough to allow participants to “see through” to its underlying statistical framework, but they argued that between-subject manipulations in nonchance settings were most informative about how people typically reasoned in everyday life.

The Positive Model: The Original Perceptual Metaphor

Along with the negative message that people do not intuitively follow Bayes’ rule, Kahneman and Tversky developed a positive descriptive model of statistical intuitions. When people infer the likelihood of a hypothesis from evidence, they asserted, people intuitively compute a feeling of certainty based on a small number of basic operations that are fundamentally different from Bayes’ rule. In particular, these basic heuristic processes include computing the similarity between a sample case and the category prototype or generating mechanism (representativeness), computing how easily instances of the relevant category come to mind (availability), and adjusting an already existing impression or number to take into account additional factors (anchoring and adjustment). Thus, representativeness measures the fit between a case and a possible cause, or between a sample and a possible distribution. Availability measures the ease with which specific examples come into consciousness: A highly unlikely event is one that seems literally “unimaginable.” Anchoring and adjustment is something quite different; it is not a measure, but a simplistic process of combination that fails to weight each component by its evidential value. These are heuristics because they are “shortcut” tools that bypass a more complicated and optimal algorithmic solution, where an algorithm is a step-by-step set of rules that guarantees a correct or optimal answer. Heuristics can be described in the language of “if then” procedural rules. *If* seeking the probability that a case is a member of a given category (or that

a sample was generated by a given population), *then* compute the similarity between the case/sample and the category/population prototype.” “*If seeking the probability that an event will occur, then compute the ease with which examples of that event come to mind.*” “*If a number is available for use and on the right scale, then adjust that number upward or downward according to knowledge that comes to mind.*” Whether such procedures were meant to be conscious or intentional strategies was not explicitly stated in the original papers.

Each of the operations described by Kahneman and Tversky yields an impression of certainty or uncertainty, but the heuristic operations themselves are unaffected by some of the required inputs to the Bayesian algorithm, such as prior odds ratio, separability of the hypotheses, validity of the evidence, or sample size. Instead, these “direct assessments” of probability are fundamentally *nonextensional* and *nonstatistical*, because they operate directly on the sample evidence without considering the relevant set-inclusion relations (the extensional rules), and without considering the degree of variability or uncertainty in the case information controlled by considerations of sample size and evidence quality (statistical rules).

In this approach, deviations from the normative model were not considered “failures of reasoning” but “cognitive illusions.” This term emphasizes that the outputs of the judgmental heuristics, like the processes involved in vision and hearing, lead to compelling impressions that do not disappear even in the presence of relevant rule-based knowledge. Furthermore, the heuristics do not represent a “strategy” chosen by the individual judge; again like perceptual processes, the heuristics produced their output without guidance or active awareness of their constructive nature. This general notion was not novel; it had been introduced by J. Cohen (1960) in his study of “psychological probability:”

Psychological probabilities which deviate from norms based on an abstract or “idealized” person are not errors, in a psychological sense, any more than optical “illusions” as such are errors. They can only be described as errors in terms of a non-psychological criterion. Knowledge of the objective lengths of the Muller-Lyler lines, for example, does not appreciably affect our subjective impressions of their magnitude. Precisely the same is true of the Monte Carlo fallacy [gambler’s fallacy] . . . ; even mathematicians who are perfectly convinced of the independence of

the outcomes of successive tosses of a coin are still inclined to predict a particular outcome just because it has not occurred for a relatively long time in a series of tosses. (Cohen, 1960, p. 29)

The heuristic approach helped to explain existing anomalies in statistical intuition as well as predict new phenomena. In particular, the gambler’s fallacy and probability matching can be seen as examples of representativeness at work. A long run is unrepresentative of a random chance process, and so we expect to see alternations to make the sequence seem more representative. In probability matching, the strategy of always predicting the most common outcome is completely unrepresentative of the kinds of patterns that seem likely to occur by chance, so predictions are made with the same kind of alternations that are representative of a random or chance process. Later, Gilovich, Vallone, and Tversky (1985) showed that people systematically misperceive random sequences because of the expectation that the sample sequence will “represent” the random nature of the causal distribution and contain many alternations and few long runs. When basketball fans were presented with a sequence of shots described as hits and misses, a majority perceived a sequence with a .5 probability of alternation as representing a “streak,” because it included more long runs than they expected. An even larger majority perceived a sequence with a .8 probability of alternation as representing a “chance” sequence, because there were few long runs, and so the observed pattern matched the defining characteristics of a “random” process. Not surprisingly, such fans perceived actual players to be streak shooters, even though none of the players studied had shooting patterns that deviated from a simple independence model based on the assumption that hits were no more likely to follow a hit than to follow a miss.

The often-observed difficulties people have in understanding and identifying regression artifacts (e.g., Campbell & Kenny, 1999) also follow from the application of representativeness: People expect an effect to be just as extreme as its cause, regardless of the strength of the predictive relationship. Thus, children are expected to be just as tall, short, or clever as their parents, and experienced psychologists expect their experimental replications to be just as significant as the original (significant) studies. Kahneman and Tversky (1973) coined the term “prediction by evaluation” to describe the process of matching the size of the effect with the size

of the cause: The extremity of the causal variable is *evaluated* and then an outcome is that is equally as extreme. However, when children are less clever than their parents or replications yield weaker results than their originals, people invariably seek out causal explanations—ignoring the statistical law of regression that operates whenever predictive relationships are not perfect. Such findings have profound implications beyond the rejection of an unrealistic model of rationality: If people see random sequences as systematic deviations from chance, and develop causal explanations for phenomena that represent simple regression artifacts, we can expect an intellectual culture that develops and maintains unfounded superstitions and useless home medical treatments, that sustains multiple competing explanations of social phenomena, and distrusts the quantitatively guided conservatism of science (Gilovich, 1991).

AVAILABILITY

Given that there are a lot of Canadian comedians, one can probably think of particular examples very readily. There is merit, then, in turning this around and concluding that if one has an easy time thinking of Canadian comedians, there probably are a lot of them. The logic is generally sound and it constitutes the essence of the *availability heuristic*, or the tendency to use the ease with which one can generate examples as a cue to category size or likelihood. But the “probably” in this inference is important. There can be other reasons why examples of a given category are easy or hard to generate and so availability is not always a reliable guide to actual frequency or probability (Kahneman & Tversky, 1973; Macleod & Campbell, 1992; Rothbart, Fulero, Jensen, Howard, & Birrell, 1978; Tversky & Kahneman, 1973).

Kahneman and Tversky (1973) first demonstrated this in a series of classic experiments. In one, participants were asked whether there are more words that begin with the letter “r” or that have “r” as the third letter. Because it’s easier to generate words that start with “r” (red, rabid, ratatouille...) than words that have an “r” in the third position (...Huron, herald, unreasonable), most participants thought there were more of the former than the latter. In reality, there are three times as many words with an “r” in the third position.

Ross and Sicoly (1979) explored the implications of the availability heuristic for everyday social life. They asked couples to specify their own percentage contribution to various tasks and outcomes

that come with living together—keeping the house clean, maintaining the social calendar, starting arguments, and so on. They predicted that each person’s own contributions would be more salient than their partner’s contributions and so both partners would overestimate their own role. When the estimates made by each member of a couple were summed, they tended to exceed the logical maximum of 100%. This was true, notably, for negative actions (e.g., starting fights) as well as positive actions—evidence that it is the availability heuristic and not self-enhancing motivations that is responsible for this effect.

Norbert Schwarz and his colleagues have shown how the availability heuristic can influence people’s self-assessments and, in so doing, also settled an important conceptual issue that lies at the core of the availability heuristic (Schwarz, Bless, et al., 1991; Schwarz & Vaughn, 2002; see also Gabrielcik & Fazio, 1984). Recall that people are assumed to use the *ease* with which they can generate instances of a given category when making judgments about the category. But note that if instances are easy to generate, one will probably come up with a lot of them. So how can we be sure that people are in fact influenced by the ease with which they generate instances (a metacognitive feature) rather than the *number* of instances they generate (a cognitive feature)? Typically, we can’t. What Schwarz and colleagues did was to disentangle these two, usually intertwined features. In one representative experiment, they asked half their participants to think of times they had been assertive and the other half to think of times they had been unassertive. Some of the participants in each group were asked to think of six examples and the others were asked to think of twelve examples. The required number of instances, six and twelve, were carefully chosen so that thinking of six examples would be easy but thinking of twelve would be a challenge.

This manipulation separates ease of generation (process) from the number of examples generated (content). Those asked to think of twelve examples of their assertiveness (or unassertiveness) will think of more examples than those asked to think of six, but they will have a harder time doing so. What Schwarz and colleagues found was that those asked to think of six examples of their past assertiveness later rated themselves as more assertive than those asked to think of twelve examples. The same pattern held for those asked to think of past examples of unassertiveness. Thus, it is the ease with which

people can recall examples, not the number of examples recalled, that dominates people's judgments. The effect was so strong, in fact, that those asked to come up with twelve examples of their own unassertiveness (and who thus had lots of examples of their failure to be assertive on the top of their heads) rated themselves as *more assertive* than those asked to come up with twelve examples of assertiveness (and who thus had lots of examples of their past assertiveness at the top of their heads.)

In a wry application of this paradigm, Fox (2006) had students list either two or ten ways a course could be improved as part of the standard end-of-the-term course evaluation process. Students asked to list ten possible improvements apparently had difficulty doing so, because they rated the course significantly more favorably than students asked to list two ways to improve.

REPRESENTATIVENESS

A university nutritionist informed readers of her column that a tomato "has four chambers and is red" and that eating tomatoes is good for the heart; a walnut "looks like a little brain" and "we now know that walnuts help develop more than three dozen neuron-transmitters (sic) for brain function;" and kidney beans assist with the healthy functioning of their organ namesake (Jones, 2008). This advice appears to be heavily influenced by a second heuristic identified by Kahneman and Tversky: representativeness.

Making judgments on the basis of representativeness reflects the mind's tendency to automatically assess the similarity between two entities under consideration and to use that assessment as input to a judgment about likelihood. Judgments about the likelihood of an object belonging to a category are powerfully influenced by how similar the object is to the category prototype (Kahneman & Tversky, 1972, 1973; Tversky & Kahneman, 1983). Judgments of the likelihood that an outcome stems from a particular cause are powerfully influenced by the similarity between putative cause and observed effect (Gilovich & Savitsky, 2002; Nisbett & Ross, 1980). Judgments about the likelihood of obtaining a given result are powerfully influenced by the similarity between the features of the imagined result and those of the processes thought to be at work (Kahneman & Tversky, 1972, 1973; Tversky & Kahneman, 1971).

The most compelling way to demonstrate that judgments are "powerfully" influenced by a hypothesized process is to show that they are excessively

influenced. Much of the research on representativeness has therefore sought to show that the heuristic leads people to make judgments that violate clear normative standards. Judging whether a sample is likely to have come from a particular generating process by assessing the similarity between the two, for example, has been shown to give rise to a "law of small numbers," or a tendency to believe, contrary to probability theory, that even small samples should be representative of the populations from which they are drawn (which is true of large samples and is captured in the law of large numbers). The belief in a law of small numbers has been established by studies showing that people (including expert statisticians and psychologists) are excessively confident about the replicability of research findings (Tversky & Kahneman, 1971), have difficulty recognizing or generating random sequences (Falk & Konold, 1997; Gilovich, Vallone, & Tversky, 1985; Wagenaar, 1972), and are overly influenced by the relative proportion of successes and failures, and insufficiently influenced by sample size, in assessments of how confident they can be in a particular hypothesis (Griffin & Tversky, 1992).

The work on representativeness that garnered the most attention and sparked the greatest controversy, however, involved experiments demonstrating that the allure of representativeness can prevent people from utilizing base rates or basic set-inclusion principles when making predictions. In one now-classic study (Kahneman & Tversky, 1973), participants were given the following description of an individual enrolled in graduate school:

Tom W. is of high intelligence, although lacking in true creativity. He has a need for order and clarity, and for neat and tidy systems in which every detail finds its appropriate place. His writing is rather dull and mechanical, occasionally enlivened by somewhat corny puns and by flashes of imagination of the sci-fi type. He has a strong drive for competence. He seems to have little feel and little sympathy for other people and does not enjoy interacting with others. Self-centered, he nonetheless has a deep moral sense.

One group of participants was asked to rank nine disciplines in terms of how closely Tom W. resembled the typical student in that field. A second group ranked them in terms of the likelihood that Tom was actually enrolled in each. A third group simply estimated the percentage of all graduate students in the United States who were enrolled in each discipline.

There were two critical findings. First, the rankings of the likelihood that Tom W. actually studied each of the disciplines were virtually identical to the rankings of how similar he seemed to the typical student in each field. Participants' assessments of likelihood, in other words, were powerfully influenced by representativeness. Second, the rankings of likelihood did not correspond at all with what the participants knew about the popularity of the different disciplines. Information about the base rate, or the *a priori* likelihood of Tom being a student in each of different fields, was simply ignored.

Experiments like this sparked a long-running controversy about whether and when people are likely to ignore or underutilize base rates (Cosmides & Tooby, 1996; Gavanski & Hui, 1992; Gigerenzer, 1991; Griffin & Buehler, 1999; Koehler, 1996). The controversy was productive, especially of publications, because it yielded such findings as: people are more likely to utilize base-rate information if it is presented after the information about the individual (Krosnick, Li, & Lehman, 1990), if the base rate is physically instantiated in a sampling paradigm (Gigerenzer, Hell, & Blank, 1988, but see Poulton, 1994, p. 153), and if the base rate is causally related to the to-be-predicted event (Ajzen, 1977; Tversky & Kahneman, 1982). But in an important respect the controversy was misguided because the essential idea being put forward was that people's judgments are powerfully influenced by representativeness, not that people never use, or even typically don't use, base rates. Instead, the Tom W. studies and others like it were existence proofs of the power of representativeness to overwhelm all other considerations in at least some circumstances.

ANCHORING

Suppose someone asks you how long it takes Venus to orbit the sun. You reply that you don't know (few people do), but your interrogator then asks for an estimate. How do you respond? You might think to yourself that Venus is closer than Earth to the sun and so it probably takes fewer than the 365 days it takes the earth to make its orbit. You might then move down from that value of 365 days and estimate that a year on Venus consists of, say, 275 days. (The correct answer is 224.7.)

To respond in this way is to use what Tversky and Kahneman called the anchoring and adjustment heuristic (Tversky & Kahneman, 1974). One starts with a salient or convenient value and adjusts to an estimate that seems right. The most notable

feature of such adjustments is that they tend to be insufficient. In most investigations of such "anchoring effects," the investigators take care to ensure that the respondents know that the anchor value is entirely arbitrary and therefore carries no implication whatsoever about what the right value might be. In the initial demonstration, Tversky and Kahneman (1974) spun a "wheel of fortune" device and then asked participants whether the percentage of African countries in the United Nations is higher or lower than the number that came up. After participants indicated whether they thought it was higher or lower, they were asked to estimate the actual percentage of African countries in the United Nations. What they found was that the transparently arbitrary anchor value significantly influenced participants' responses. Those who confronted larger numbers from the wheel of fortune gave significantly higher estimates than those who confronted lower numbers. Anchoring effects using paradigms like this have been observed in people's evaluation of gambles (Carlson, 1990; Chapman & Johnson, 1999), estimates of risk and uncertainty (Plous, 1989; Wright & Anderson, 1989), perceptions of self-efficacy (Cervone & Peake, 1986), anticipations of future performance (Switzer & Sniezak, 1991), answers to general knowledge questions (e.g., Jacobowitz & Kahneman, 1995), and willingness to pay for consumer items (Ariely, Loewenstein, & Prelec, 2003).

As the research on anchoring evolved, comparable effects using all sorts of other paradigms have been observed and it appears that such effects are not always the result of insufficient adjustment. Indeed, probably the fairest reading of the anchoring literature is that there is not one anchoring effect produced by insufficient adjustment, but a family of anchoring effects produced by at least three distinct types of psychological processes (Epley, 2004). Epley and Gilovich (2001, 2004, 2005, 2006) have provided evidence that people do indeed adjust insufficiently from at least some anchor values, particularly those that people generate themselves (like the earlier question about Venus). They have found, for example, that people articulate a process of adjusting from self-generated anchors, and that manipulations that should influence adjustment, but not other potential causes of anchoring, have a significant effect on people's judgments. In particular, people who are incidentally nodding their heads while answering, are cognitively busy, or lack incentives for accurate responding tend to be more influenced by *self-generated* anchor values than those

who are incidentally shaking their heads, are not busy, or are given incentives for accuracy.

Manipulations such as these, however, have generally been shown to have no effect on participants' responses in the standard (experimenter-generated) anchoring paradigm pioneered by Tversky and Kahneman (Chapman & Johnson, 1999; Epley & Gilovich, 2001, 2005; Tversky & Kahneman, 1974; see Simmons, Leboeuf, & Nelson, 2010 for an exception). At first glance, this is a bit of a puzzle because it raises the question of why, without insufficient adjustment, anchoring effects would occur. This question has been addressed most extensively by Thomas Mussweiler and Fritz Strack (Mussweiler, 2002; Mussweiler & Strack, 1999, 2000; Strack & Mussweiler, 1997). They maintain that most anchoring effects are the result of the enhanced accessibility of anchor-consistent information. The attempt to answer the initial question posed by the investigator—"Is the Nile longer or shorter than 5,000 [800] miles?"—leads the individual to first test whether the given value is correct—is the Nile 5,000 [or 800] miles long? Because people evaluate hypotheses by attempting to confirm them (Evans, 2007; Snyder & Swann, 1978; Skov & Sherman, 1986), such a search generates evidence disproportionately consistent with the anchor. Mussweiler and Strack (2000) provide support for their analysis by showing that information consistent with the anchor value presented to participants is indeed disproportionately accessible. For example, participants who were asked whether the price of an average German car is higher or lower than a high value were subsequently quick to recognize words associated with expensive cars (Mercedes, BMW); those asked whether the price of an average German car is higher or lower than a modest value were subsequently quick to recognize words associated with inexpensive cars (Volkswagen, Golf).

Oppenheimer, LeBoeuf, and Brewer (2008) have recently shown that the semantic activation elicited by different anchors can be quite general. They asked one group of participants whether the Mississippi River was longer or shorter than 4,800 miles, and another group whether it was longer or shorter than 15 miles. They then asked their participants to draw a line equal to the length of a standard toothpick. Those exposed to the high initial anchor drew longer toothpicks than those exposed to the low initial anchor. This suggests that exposure to the initial anchor activated the general concept of "long" or "short," which influenced their representation (and

production) of a standard toothpick. To test this idea, Oppenheimer and colleagues had participants in a follow-up experiment perform a word completion task after being exposed to high or low anchor values. Participants exposed to the high anchors were more likely to form words connoting bigness (BIG for B_G, LONG for _ONG) than those exposed to the low anchors.

Recent research suggests that there is likely a third source of anchoring effects: pure numeric priming. That is, an anchor activates its own numeric value and those close to it, which are then highly accessible and influential when the person tries to fashion a response. In one notable experiment, participants were asked whether the runway at Hong Kong International Airport was longer or shorter than 7.3 kilometers or 7,300 meters and were then asked to estimate the cost of an unrelated project. Those asked the question in terms of meters gave higher estimates on the second, unrelated task than those asked the question in terms of kilometers—presumably because the latter primed smaller absolute numbers (Wong & Kwong, 2000). Although some have argued otherwise, this does not appear to be the result of the differential accessibility of semantic information consistent with the initial anchor because 7.3 kilometers and 7,300 meters represent the same value, just in different units. More recent research casts further doubt on the possibility that the differential accessibility of anchor-consistent semantic information is responsible for such effects. Critcher and Gilovich (2008) asked participants what percentage of the sales of a P-97 (or P-17) cell phone would be in the European market. Participants estimated a higher percentage of European sales for the P-97 than the P-17. Note that the process that would give rise to the heightened accessibility anchor-consistent semantic information (testing whether the anchor value might be the correct value) is not applicable here. It seems far-fetched to maintain that participants asked themselves whether part of the model label (97 or 17) might be the European market share.

The Negative Model: Normative Neglect and Its Discontents

The "negative" conclusion from this program of research—that intuitive judgments typically reflect only case-specific evidence; that people neglect base rates, evidence diagnosticity, sample size, and other features about the broader distribution—is enough to explain many of the anomalies in probability

judgment listed earlier. If people focus only on the sample-specific evidence, then conservatism should be prevalent when base rates, sample sizes, evidence, and diagnosticity are high, and radical or overconfident judgments should prevail when they are low. This “psychology of evidential neglect” was implicit in the defining papers in the H&B program, and it was later made explicit by Griffin and Tversky’s (1991) “strength-weight” theory and then modeled by Brenner’s (1995, 2003) random support theory. Koehler, Brenner, and Griffin (2002) found substantial support for the basic neglect model in the everyday probabilistic judgments of physicians, economists, and lawyers working in real-world settings. Even weather forecasters, aided by computer projections and immediate outcome feedback, showed substantial neglect of base rate and validity considerations until they received specific feedback about their biases.

As noted earlier in the discussion of the base-rate fallacy and the Tom W. study in particular, criticisms of the “neglect” message began soon after the early laboratory studies were published. One prominent critic claimed that he had “disproved the representativeness heuristic almost before it was published; and therewith...also disproved the base rate fallacy” (Anderson, 1996, p. 17). In particular, Anderson had shown that base rates and case-specific information received about equal weight when manipulated across scenarios in a within-subject design. Tversky and Kahneman (1982) accepted that within-subject designs revealed the *capacity* for rule-based thinking, whereas between-subject designs revealed the *actual application* of rules in practice.

Many economists, whose theories would suffer most if the H&B challenge to classical rationality was widely accepted, wondered about whether the various neglect biases would disappear with appropriate incentives or market conditions. In a series of studies, an economist (Grether, 1992) found that judgments consistent with the Bayesian model did increase very slightly, but significantly, with incentives for accuracy. More important, even in a chance setup (balls sampled from bingo cages), with both sample evidence and base rates determined by drawing balls from a cage, a context that should make the sampling mechanism salient and “transparent,” there was still considerable evidence of heuristic thinking. Similarly, studies of business students in market games involving repeated plays and real incentives also revealed biased judgments in accord with the H&B account (Camerer, 1987), but the biases

seemed to decline with repeated playing of the game (see Camerer & Smith, Chapter 18). It is important to note, however, that studies of judgment in which people actively discovered the base rate for themselves (instead of deciding which of the experimenter’s numbers was relevant to the task) also support a strong form of base-rate neglect (e.g., Dawes, Mirels, Gold, & Donahue, 1993; Griffin & Buehler, 1999; Yates & Estin, 1996; Yates et al., 1998).

We next turn to two critiques that have attracted considerable research attention and raise questions about the fundamental underpinnings of the first wave of the H&B research. The first claim is that findings of the program are merely artifacts of the conversational rules between subject and experimenter; the second is that H&B researchers have confused different definitions of probability.

Some commentators have claimed that many of the apparently “irrational” judgments observed in various studies were actually caused by rules of conversational implicature. There are two versions of this claim: The first is that people actively make sense of their environment, actively search for the appropriate meaning of questions, statements, conversations, and questionnaires, and that the same objective information can mean something different in different social or conversational contexts. This perspective is part of a *constructivist* approach to judgment (Griffin & Ross, 1991) that is consistent with the second wave of H&B research discussed later (e.g., Kahneman & Miller, 1986). Kahneman and Tversky (1982b) themselves discussed the problems with using what they called the “conversational paradigm” and noted that participants were actively involved in figuring out what the experimenters wanted to convey, just as if they were engaged in a face-to-face conversation. In that same chapter, they further noted the relevance of Grice’s maxims of communication to their problems (see Grice, 1975; Hilton & Slugoski, 2000) and later explicitly attempted to develop judgment tasks that avoided the common-language ambiguity of terms such as “and” and “or” (Tversky & Kahneman, 1983).

However, this acknowledgment did not prevent a second and more critical version of the conversational interpretation. The claim is that results of the scenario studies lacked external validity because changes in wording and context could reduce the rate of biased responses to questionnaire scenarios. For example, Macchi (1995) argued that base-rate neglect may arise from textual ambiguity such that the verbal expression of $P(D|H_1)$ is interpreted as

$P(H1|D)$. Thus, the text “The percentage of deaths by suicide is three times higher among single individuals . . .” may be interpreted to mean “within the suicide group the percentage of single individuals who died by suicide is three times higher” (p. 198). To test this hypothesis, Macchi changed the key phrase to read “1% of married individuals and 3% of single individuals commit suicide” and found that this dramatically increased the number of participants who used both the base rate and the specific information provided. Of course, it is possible to reapply a conversational analysis to the revised question, and it is difficult to know when the cycle should end. That is why it is so useful to have a real-world phenomenon to guide the evaluation of laboratory studies that otherwise can get lost in a perpetual cycle of “experiments about experiments.”

The conversational perspective has also focused on the lawyer-engineer paradigm. Some follow-up studies challenged the explanation that the base-rate neglect observed in the original paradigm was due to judgment by representativeness, and they have been widely cited as evidence that heuristic thinking is eliminated in familiar, real-world social settings (e.g., Barone, Maddux, & Snyder, 1997). For example, Zukier and Pepitone (1984) found greater attention to base-rate information when participants were instructed to think like scientists than when participants were instructed to understand the person’s personality. A related study (Schwarz, Strack, Hilton, & Naderer, 1991) reported greater attention to base-rate information when participants were told that the personality sketch was randomly sampled by a computer than when they were told it was written by a psychologist. Thus, one might be tempted (despite the many other demonstrations of representativeness in the laboratory and the real world) to conclude that the proper use of statistical logic depends largely on social roles and contextual implications. However, a closer look at these studies leads to an interpretation more in line with a “constructive” sense-making interpretation that does not undercut the H&B position. In both studies participants were presented only with a low base rate of engineers; inferences about base-rate use were based on changes in judgment in a paradigm that did not manipulate base rate. Thus, these studies suggest not that judgment by representativeness is an artifact of a contrived experimental situation, but rather that heuristics operate upon information that is actively constructed by the perceiver. As noted, one of the first examinations of the role of conversational

implicature (Grice, 1975) in the H&B paradigm was by Kahneman and Tversky in the final chapter of the Kahneman, Slovic, and Tversky (1982) book (Kahneman & Tversky, 1982b, p. 502), which marks the boundary between the first and second waves of the H&B research tradition.

The second major critique is the claim that the H&B program is built solely (and narrowly) on questions about probability judgments for unique events. Some defenders of the “objective” or “frequentist” school of probability have denied any role for the rules of probability in describing events that cannot be replicated in an infinite series. Nonetheless, it is undeniable that physicians, judges, and stockbrokers, along with virtually everyone else, use terms such as “probability” and “chance” to describe their beliefs about unique events. One of the greatest statisticians of the 20th century has described the logical foundation of the subjective probability viewpoint as follows: “the formal rules normally used in probability calculations are also valid, as conditions of consistency for subjective probabilities. You must obey them, not because of any logical, empirical or metaphysical meaning of probability, but simply to avoid throwing money away” (De Finetti, 1970). We note that this point can also be made with respect to throwing away lives, or even throwing away happiness.

The frequentist critics of the H&B approach claim that when the classic demonstrations of heuristics are reframed in terms of aggregate frequency, the biases decline substantially or even disappear (e.g., Cosmides & Tooby, 1996; Gigerenzer, 1994; 1998; Jones, Jones, & Frisch, 1995). However, proponents of the H&B approach have explored this possibility for some time, in what we term the “second wave” of heuristics research. For example, Kahneman and Tversky (1982b) proposed that when making aggregate frequency judgments, people were more likely to recruit statistical rules of reasoning, especially rules of set-inclusion relationships, than when making individual probability judgments; Tversky and Kahneman (1983) proposed that set-inclusion relations were more compelling arguments when framed in frequentistic “counting” contexts; Griffin and Tversky (1992) proposed that aggregate frequency judgments led to greater attention to “background” information such as past performance (including base rates); and Tversky and Koehler (1994) proposed that the violations of set-inclusion relations observed when compound hypotheses were explicitly “unpacked” into elementary

hypotheses would be smaller for frequency than probability judgments. Thus, the dispute between critics and proponents of the H&B tradition is not about whether probability and frequency judgments are psychologically distinct, or that frequency presentations are intrinsically simpler than probability interpretations, or even that the magnitude of biases are typically smaller in frequentistic formulations—the dispute is about the causes of the discrepancy and its implications for understanding the classic demonstrations of judgmental heuristics and heuristic thinking in real-world applications.

According to the H&B approach, the discrepancy between single-event probability and aggregate frequency judgments occurs because aggregate frequency judgments are less amenable to heuristic assessments that operate “holistically” on unique cases and are more sensitive to statistical or logical rules because the application of such rules is more transparent. Furthermore, comparisons of the two tasks involve irrelevant confounds because the two scales of judgment are rarely psychologically parallel (Griffin & Buehler, 1999). According to H&B’s frequentist critics, a “frequency format” is consistent with the evolved software of the mind, and single-event “subjective” probability judgments are inherently unnatural (Gigerenzer, 1998, 1994). Supporting this perspective is evidence that people are extremely efficient, and seemingly unbiased, at encoding and storing the frequencies of letters and words to which they have been exposed. On the other hand, this perspective cannot account for the observation that virtually all uses of the concept “chance” (meaning likelihood) in early English literature are consistent with a subjective, single-event judgment (Bellhouse & Franklin, 1997), nor that people untutored in Bayesian or frequentist statistics regularly use expressions of subjective probability to describe their beliefs about the world. A series of studies by Sloman and his colleagues has provided convincing evidence that frequentistic representations improve probability judgments when and if they lead to more concrete representations of set-inclusion relations (e.g., Barbey & Sloman, 2007; Sloman, Over, Slovak, & Stiebel, 2003).

Heuristics Unbound: Beyond Three Heuristics

As with any initial statement of a theory, the first wave of H&B demonstrations left some empirical anomalies to be explained. One prominent issue was the problem of “causal base rates” (Ajzen, 1977):

When base rates could be given a causal interpretation (e.g., a high proportion of failures on an exam implied that a difficult exam *caused* the failure rate), they received substantial weight in judgment. This led Tversky and Kahneman (1982) to include the computation or assessment of causality or causal propensity (see Cheng & Buehner, Chapter 12) as a basic heuristic operation, and to acknowledge that the distinction between case-specific and population-based information was less sharp than originally proposed. This latter conclusion was reinforced by the finding that people were sometimes most responsive to the size of a sample *relative* to the size of a population (Bar-Hillel, 1982). Such a “matching” approach to sample size implied a broader kind of representativeness calculation, or as Bar-Hillel termed it, a second-order representativeness. The sharp distinction between heuristics that operated on cases, and rules that operated on abstract statistical quantities, it appeared, was not always clear and seemed better captured by a more flexible distinction between “holistic” and “analytic” thinking. Furthermore, the initial statements of the H&B approach contained some ambiguity with regard to whether judgmental heuristics were deliberate strategies to avoid mental effort or were largely automatic processes that were uncontrolled and uncontrollable. These issues were addressed by a second generation of papers on judgmental heuristics by Kahneman and Tversky.

The second wave of heuristics research began with an analysis of the “planning fallacy,” the tendency for people to make optimistic predictions even when aware that similar projects have run well over schedule (Kahneman & Tversky, 1979). This paper introduced a new perceptual metaphor, based on prediction by evaluation, that contrasted an “inside” and an “outside” perspective on a prediction problem. Using an inside or internal perspective, a judge focuses on the specific details of the current case; using an outside perspective, a judge “sees” the specific case as one instance of a broader set of instances. Shortly afterward, a paper on causal reasoning (Tversky & Kahneman, 1982) demonstrated how intuitive or heuristic processes could be applied to both case-specific and distributional information as long as both types of information were in a form amenable to “natural assessments.” For example, base rates that have causal implications (e.g., a sports team that has won 9 of its last 10 games) may induce a computation of a “causal disposition” connected to that team (Kahneman & Varey, 1990a). These two

approaches blurred the distinction between case-specific and statistical information, and instead distinguished between information that can be directly evaluated by natural assessments in a holistic manner (“associationist” computations) and information that requires logical inference (rule-based computations) before it can be used.

A key paper in this second wave of research included the exploration of the *conjunction fallacy* (Tversky & Kahneman, 1983). Although cited primarily for the memorable “Linda problem,” the 1983 paper on the conjunction fallacy further developed the perceptual model of judgmental heuristics and clarified the role of abstract rules in intuitive statistical judgment. In this and related papers (e.g., Kahneman & Miller, 1986; Kahneman & Tversky, 1982a), Kahneman, Tversky, and colleagues distinguished low-level “natural” or routine or basic cognitive assessments that are relatively automatic and spontaneously evoked by the environment, from explicit, higher level judgmental heuristics, which are typically evoked by an attempt to answer a question. Clear candidates for natural assessments include computations of similarity, causal potency, and counterfactual surprise.

Tversky and Kahneman (1983) chose the conjunction rule of probability as a case study in the conflict between heuristic thinking and rule-based reasoning. They argued that the conjunction rule of probability (no conjunction of events can be more probable than either constituent event alone) is one of the most basic and compelling rule of probability and is understood, in some form, by virtually every adult. Thus, in a wide variety of contexts, they examined when the *conjunction rule* would overcome the “conjunction fallacy,” the tendency to judge a conjunction as more probable than its least likely constituent.

For example, participants in one study were given the following description of an individual:

Bill is 34 years old. He is intelligent but unimaginative, compulsive, and generally lifeless. In school, he was strong in mathematics but weak in social studies and humanities.

They were then asked to rank the likelihood of eight possible life outcomes for Bill, including (1) Bill is an accountant, (2) Bill plays jazz for a hobby, and (3) Bill is an accountant who plays jazz for a hobby. Ninety-two percent of the respondents assigned a higher rank to (3) than to (2), even though any state of the world that satisfies (3) automatically satisfies (2) and so (3) cannot be more likely than (2).

Because the conjunction fallacy violates one of the most basic rules of probability theory, Kahneman and Tversky (1983) anticipated controversy and provided a wide-ranging discussion of alternative interpretations. They included additional controls for the possibility that respondents misunderstood the words “and” or “or;” they made sure that the same effects occurred with frequencies as well as probabilities and that the effect applied when reasoning about heart attacks as well as when reasoning about personality descriptions; and they made sure that the same effects obtained with expert and seasoned political forecasters as with college students. Nonetheless, the anticipated controversy ensued, centering around participants’ interpretation of the conjunction (e.g., Mellers, Hertwig, & Kahneman, 2001), the effects of frequency versus probability response formats (Hertwig & Gigerenzer, 1999), and the limits of laboratory research.

Kahneman and Tversky created conjunctions that “seemed” or “felt” more likely than their constituents by using representativeness (combined events or descriptions were more similar to the target than one or both of the constituents, as in the Bill and Linda examples), availability (the combination of events or descriptions were better search cues than one or both of the constituents), and causal relatedness (the combination of events created a causal link that seemed plausible, easy to imagine, and therefore more likely than one or both of the constituent events). The real-world phenomenon that is reflected in the findings from the conjunction fallacy studies is that as predictive scenarios become more detailed, they become objectively more unlikely yet “feel” more likely. The authors noted that many participants reported being simultaneously aware of the relevance of the conjunction rule and the feeling that the conjunction was more likely than the constituent categories. Conjunction fallacies were extremely common in between-subject designs, quite common in nontransparent within-subject designs, and only substantially reduced by a combination of a within-subject design and a frequentistic design in which participants could “see” that the number of people with A *and* B must be less than the number of people with A. Except in special circumstances, then, intuitive judgments do not conform to the rules of probability, even when those rules are known and endorsed by the intuitive judges.

Note how this notion is fundamentally different from the “cognitive miser” model of social cognition. Heuristic judgments are not explained as the

result of too little thought due to cognitive laziness or inadequate motivation, but as the result of uncontrolled “thinking too much” in quick and natural ways. This model of spendthrift automatic processes was termed “mental contamination” by Kahneman and Varey (1990b), who related the basic processes of heuristic thinking to a wide range of perceptual, cognitive, and social examples, including the Stroop effect and motor effects on persuasion.

Whereas the original H&B program focused on situations in which only heuristics were evoked, and the conjunction fallacy paper examined how heuristics and statistical rules might compete, Griffin and Tversky (1991) described how the strength of impression and the weight of statistical evidence might combine. Using the anchoring and adjustment process as the “master heuristic,” they suggested that people typically anchor on the strength of their impressions and then adjust (insufficiently) according to rule-based arguments about sample sizes or evidential validity. In “support theory,” Tversky and his students developed a formal treatment of how perceptions of evidential support are translated into judgments of probability.

Support theory (Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994) was founded, in particular, on earlier observations of systematic violations of extensionality such as the conjunction fallacy and findings from the fault-tree paradigm (Fischhoff, Slovic, & Lichtenstein, 1978). In contrast to probability theory, in which probabilities are assigned to events that obey the laws of set inclusion, in support theory probabilities are assigned to descriptions of events, referred to as *hypotheses*. Support theory thereby allows two different descriptions of the same event to receive different probability judgments, in much the same way that prospect theory accommodated the possibility that different choices might be made to an identical decision depending on how that decision is “framed” or described (Kahneman & Tversky, 1979).

Support theory represents judged probability in terms of the balance of perceived evidential support for a focal hypothesis A and an alternative hypothesis B, such that $P(A,B) = s(A) / [s(A) + s(B)]$. For instance, A might represent the hypothesis that Jack is a lawyer and B the hypothesis that Jack is an engineer. A key feature of support theory is the assumption that hypotheses described at a greater level of detail will tend to have greater perceived support than a hypothesis describing the same event in less detail. For instance, “unpacking” the hypothesis

that Jack is a lawyer into the hypothesis that Jack is either a corporate lawyer, a criminal lawyer, a divorce lawyer, or a tax lawyer tends to increase support and hence, in the earlier example, the judged probability that Jack is a lawyer rather than an engineer. Unpacking the engineer hypothesis, by contrast, is expected to decrease the judged probability that Jack is a lawyer rather than an engineer. Such unpacking effects are particularly likely when the unpacked components are plausible but unlikely to come to mind in evaluating the packed version of the hypothesis; by contrast, unpacking implausible components can actually have the opposite effect, making the unpacked hypothesis seem less likely than its packed counterpart (Sloman, Rottenstreich, Wisniewski, Hadjichristidis, & Fox, 2004).

Support theory offered an overarching account of intuitive probability judgment that could accommodate a variety of heuristic and other reasoning processes by which the judge might evaluate the extent to which a hypothesis is supported by the available evidence. For instance, given a personality sketch of Jack, support for the lawyer hypothesis might be evaluated based on representativeness (i.e., his similarity to a prototypical lawyer). Or, in the absence of personality information, support might be based on the availability of male lawyers in memory. Support for the lawyer hypothesis might even be based on overall frequency or base-rate information. It is notable that the means by which such statistical information is incorporated in the assessment of support need not necessarily follow the specific combination formula of Bayes rule, but it simply serves as an additional argument that may be considered as part of the support assessment process.

In short, support theory continued the development of incorporating a broader set of assessment processes that went beyond basic heuristics and also characterized the role that heuristics play in intuitive probability judgment as a means by which people evaluate how much a body of evidence supports a particular hypothesis. Indeed, many studies using support theory have shown that eliciting ratings of heuristic attributes such as perceived similarity or causal strength and using them as a proxy measure of support in the earlier support theory equation can reproduce intuitive probability judgments quite closely (e.g., Fox, 1999; Koehler, 1996; Tversky & Koehler, 1994). More generally, in terms of the distinction made by Griffin and Tversky, such research has revealed that perceived support typically is highly sensitive to the strength of the available

evidence and largely insensitive to its weight or credence (e.g., Brenner, Griffin, & Koehler, 2012). This suggests that support is often evaluated in a heuristic manner, though the support theory framework itself can accommodate other, or additional, considerations as well.

Along with these developments, led by Kahneman and Tversky, the second wave of research on judgmental heuristics saw substantial contributions from other cognitive and social psychologists. Two notable extensions included the development of the “affect heuristic” by Paul Slovic and colleagues, and the splitting off of perceptual fluency from the availability heuristic and treating it as an additional metacognitive “natural assessment” along with ease of generation. The affect heuristic uses one’s immediate good/bad affective reactions to stimuli as an input to various judgments and decisions such as valuation, agreement, and more generally, approach and avoidance (Slovic, Finucane, Peters, & MacGregor, 2002).

AVAILABILITY’S CLOSE COUSIN: FLUENCY

The mere act of imagining an outcome can make it seem more likely to occur. Imagining one candidate winning an election makes it seem more likely that that candidate will triumph (Carroll, 1978) and imagining what it would be like to have a disease makes it seem that one is more at risk of getting it (Sherman, Cialdini, Schwartzman, & Reynolds, 1985). This effect was originally interpreted as the result of availability: Imagining the event made it more cognitively available and hence it was judged more likely. But what exactly are the “relevant instances” that easily (or not) come to mind when one is asked to estimate the likelihood of having an ulcer?

Another interpretation of these findings centers around the concept of fluency: Thinking of a target event is likely to have a different feel if one had, in fact, mentally tried it on earlier. It is likely to feel more “fluent.” Fluency refers to the experience of ease or difficulty associated with perception or information processing and is somewhat distinct from the ease of generating instances. A clear image is easy to process and fluent. A phonemically irregular word is hard to process and disfluent. People use the metacognitive experience of fluency as a cue when making inferences about all sorts of judgments (Jacoby & Dallas, 1981; Oppenheimer, 2008). People judge fluent names to be more famous (Jacoby, Woloshyn, & Kelley, 1989), fluent objects to be better category members (Whittlesea & Leboe, 2000), and adages

that rhyme to be more valid than those that don’t (McGlone & Tofaghbakhsh, 2000).

In addition to these direct effects on judgment, fluency appears to influence how people process relevant information. A feeling of disfluency while processing information appears to undermine people’s confidence in what they are doing, leading to something of a “go slow, be careful” approach to judgment and decision making. Thus, people are more likely to choose a default option when choosing between consumer products that are made disfluent (Novemsky, Dhar, Schwarz, & Simonson, 2007). Fluency also appears to influence the level of abstraction at which information is encoded. Given that blurry (disfluent) objects tend to appear to be farther away than distinct objects (Tversky & Kahneman, 1974), one might expect disfluent entities more generally to appear relatively far away. Indeed, Alter and Oppenheimer (2008) found that cities are judged to be farther away when their names are presented in a difficult-to-read font.

A Third Wave: Dual-Process and Two-System Accounts of Judgment Heuristics

As Neisser (1963) noted in an early review of dual modes of cognition, “The psychology of thinking seems to breed dichotomies.” Consistent with this observation, social and cognitive psychologists have recognized that people appear to approach various cognitive tasks in two very different ways (Chaiken & Trope, 1999; Evans, 2004; Kahneman, 2011; Sloman, 1996; Strack & Deutsch, 2004). One involves mental processes that are fast, associationist, and often automatic and uncontrolled. The other involves processes that are slower, rule based, and more deliberate. Scholars in both disciplines have devoted a lot of energy trying to specify the nature of these two types of processes, or “systems” of thought, and to delineate when each is operative and how they interact when people make a judgment or choose a course of action. The two systems have been given many names and, following Stanovich (1999), we refer to them simply as “System 1” and “System 2” for ease of exposition (see Evans, Chapter 8; Stanovich, Chapter 22). (To our knowledge, the term “dual processes” first appeared in Wason & Evans, 1975.)

Given Daniel Kahneman’s long-standing interest in visual attention—instantiated in his classic 1973 book *Attention and Effort*—it is not surprising that the H&B program came to incorporate

both controlled and automatic processes. (In the preface to his 1973 book, Kahneman wrote: “While the allocation of attention is flexible and highly responsive to the intentions of the moment, there are pre-attentive mechanisms that operate automatically, outside voluntary control...it is easy to notice several aspects or attributes of an object, but it is difficult or impossible to prevent the perceptual analysis of irrelevant attributes” [p. 7].) The 1983 paper on the conjunction fallacy implicitly provided a dual-system analysis of the competition between automatic intuitive processes and effortful rule-based processes, and this analysis was formalized by Sloman (1996) who also drew upon social psychological models of the conflict between a “gut feeling” and a more considered analysis (Denes-Raj & Epstein, 1994; Epstein, 1991).

Perhaps the most striking evidence of two mental systems that guide judgment and behavior is Epstein’s work on the “ratio bias” phenomenon (Denes-Raj & Epstein, 1994; Epstein, 1991). Epstein told participants that they could win a prize by blindly selecting a jellybean of a given color from one of two urns. One urn had 1 winning jellybean and nine of another, losing color. The second urn had 9 winning jellybeans and 91 of the losing color. What Epstein found was that many participants chose to select from the larger urn that offered lower odds of winning because they couldn’t resist the thought that the larger urn had more winning beans. They did so despite the fact that the chances of winning with each of the urns was explicitly provided for them. When the choice was between a 10% chance in the small urn and a 9% chance in the large urn, 61% of the participants chose the large urn. When it was a contest between 10% in the small urn and 5% in the large urn—odds only half as good in the latter—23% of the participants still chose the large urn.

Epstein attributes this decidedly irrational result to an “experiential” system of reasoning that operates on concrete representations, and hence finds the greater number of winning jellybeans in the large urn to be more promising. This experiential or intuitive impulse, however, usually conflicts with the rational realization that the actual odds are better in the small urn. Some participants explicitly stated that they knew they should pick the smaller urn, but they nonetheless were going with a gut feeling that they were more likely to pick a winner from the large one. This experience of being pulled in two different directions suggests that there are two things—two mental systems—doing the pulling.

This was emphasized by Sloman (1996), who described a possible cognitive architecture consisting of two relatively independent systems to explain the diverse findings implicating dual processes in reasoning, choice, and judgment.

Kahneman and Frederick (2002, 2005; Kahneman, 2011) highlighted these relations between System 1 and System 2 in their influential “third wave” restatement of the H&B program of research. In their “attribute substitution” account, System 1 automatically computes an assessment with some connection to the task at hand—an emotional reaction, a sense of fluency, the similarity between examples or between an example and a category. Both the perceived relevance of the assessment and its immediacy often give rise to the sense that the task is done and that the assessment produced by System 1 *is* the answer being sought. For example, one cause of death is judged to be more common than another because it is easier to think of examples of the former (Slovic, Fischhoff, & Lichtenstein, 1982). One attribute (ease of retrieval) *substitutes* for another, desired attribute (likelihood).

In many circumstances, however, and for a variety of different reasons, System 2 intervenes and deems the automatic assessment inadequate for the task at hand. A more deliberate, rule-based response is given. For example, one might consciously realize, especially if one has received training in statistics and recognizes threats to validity, that a given cause of death is highly available because it is frequently discussed in the media. “In the context of a dual-system view, errors of intuitive judgment raise two questions: ‘What features of system 1 created the error?’ and ‘Why was the error not detected and corrected by system 2?’” (Kahneman & Frederick, 2005, p. 268). The attribute substitution model has captured a great deal of attention because it offered a unified account of a diverse set of judgmental phenomena, such as the role of heuristics and logical rules on probability judgment, happiness assessments, duration neglect in remembered pain, and on contingent valuation methods used to assess people’s willingness to pay (WTP) for such things as environmental remediation. As Kahneman and Frederick (2005, p. 287) summarized:

The original goal of the heuristics and biases program was to understand intuitive judgment under uncertainty. Heuristics were described as a collection of disparate cognitive procedures, related only by their common function in a particular judgmental

domain... It now appears, however, that judgment heuristics are applied in a wide variety of domains and share a common process of *attribute substitution* in which difficult judgments are made by substituting conceptually or semantically related assessments that are simpler and more readily accessible.

The current treatment explicitly addresses the conditions under which intuitive judgments are modified or overridden. Although attribute substitution provides an initial input into many judgments, it need not be the sole basis for them. Initial impressions are often supplemented, moderated or overridden by other considerations, including the recognition of relevant logical rules and the deliberate execution of learned algorithms. The role of these supplemental or alternative inputs depends on characteristics of the judge and the judgment task.

Although the dual-system account of judgmental heuristics is not without its skeptics (e.g., Keren & Schul, 2009), in its very general form it has received broad acceptance. A number of questions, remain, however, about the best way to characterize the operations and interactions of the two systems. Evans (e.g., 2008; see Evans, Chapter 8), for instance, distinguishes default-interventionist dual-process models from parallel-competitive dual-process models. It has been suggested, furthermore, that System 2 should be split into two components reflecting cognitive ability and thinking dispositions, respectively, yielding a triprocess model (Stanovich, 2009). In short, the dual-system approach to judgment under uncertainty has been very influential, but many details will need to be filled in before it can be developed into a comprehensive process-based account of how heuristics operate.

Conclusions

We provided a historical overview of the H&B tradition, its intellectual forebearers, and its evolution through three waves of conceptualization and reconceptualization, but this should not be taken to imply that the program is frozen in the past. In addition to the hearty oak tree of classic H&B research, the program still continues to send out new green shoots of intellectual offspring. One “second-wave” example is a new model of counterfactual reasoning based on a model of semantic evidence that follows from the H&B approach (Miyamoto, Gonzalez, & Tu, 1995). The dual-processing perspective has motivated neuropsychological studies attempting to isolate and locate the brain networks associated with

the “dueling” heuristic and rule-based processes underlying classic H&B demonstrations (e.g., De Neys & Goel, 2011); the impact of experienced versus presented statistical information continues to be an active area of research (e.g., Brenner, Griffin, & Koehler, 2005; Brenner, Griffin, & Koehler, 2012; Fox & Hadar, 2006; Hertwig, Barron, Weber, & Erev, 2004); and the applied impact of the H&B tradition on understanding expert judgment in such fields as finance, political science, law, medicine, and organizational behavior continues to grow (e.g., Koehler, Brenner, & Griffin, 2002; Tetlock, 2005).

Future Directions

Questions to guide future development:

What is the relation between general cognitive ability and susceptibility to heuristic-based judgmental bias?

Are there fundamental individual differences in the tendency to make heuristic-based judgments?

What circumstances facilitate detection and correction of conflict between heuristic and rule-based evaluations?

Are the processes underlying System 2 operations that support “override” of initial, heuristic responses related to more basic inhibitory operations that guide attention in, for example, Stroop, flanker, and go/no-go tasks?

Under what conditions does extended experience in carrying out a particular judgment task reduce susceptibility to base-rate neglect, conjunction errors, and other systematic biases?

Do organizational practices or market interaction consistently attenuate biases associated with use of judgmental heuristics?

Can anchoring and/or adjustment usefully be viewed as the “master heuristic?”

How useful is the distinction between positive versus negative contributions in theory development to psychology more generally?

Acknowledgments

This chapter draws extensively upon reviews presented in Gilovich and Griffin (2010) and Griffin, Gonzalez, and Varey (2001). We acknowledge financial support from the Social Sciences and Humanities Research Council of Canada (SSHRC, Griffin), the Natural Sciences and Engineering Research Council of Canada (NSERC, Koehler), and the National Science Foundation (NSF, Gonzalez, Gilovich).

References

- Adams, J. K., & Adams P. A. (1961). Realism in confidence judgments. *Psychological Review*, 68, 33–45.

- Ajzen, I. (1977). Intuitive theories of events and the effects of base-rate information on prediction. *Journal of Personality and Social Psychology*, 35, 303–314.
- Alter, A. L., & Oppenheimer, D. M. (2008). Effects of fluency on psychological distance and mental construal (or why New York is a large city, but *New York* is a civilized jungle). *Psychological Science*, 19, 161–167.
- Anderson, N. H. (1996). Cognitive algebra versus representativeness heuristic. *Behavioral and Brain Sciences*, 19, 17.
- Ariley, D., Loewenstein, G., & Prelec, D. (2003). “Coherent arbitrariness”: Stable demand curves without stable preferences. *The Quarterly Journal of Economics*, 118, 73–105.
- Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Brain and Behavioral Sciences*, 30, 241–298.
- Bar-Hillel, M., (1982). Studies of representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 69–83). New York: Cambridge University Press.
- Barone, D. F., Maddux, J. E., & Snyder, C. E. (1997). *Social cognitive psychology*. New York: Plenum Press.
- Bellhouse, D. R., & Franklin, J. (1997). The language of chance. *International Statistical Review*, 65, 73–85.
- Brenner, L. A. (1995). A stochastic model of the calibration of subjective probabilities. Unpublished Ph.D. dissertation, Stanford University, Palo Alto, CA.
- Brenner, L. A. (2003). A random support model of the calibration of subjective probabilities. *Organizational Behavior and Human Decision Processes*, 90, 87–110.
- Brenner, L., Griffin, D. W., & Koehler, D. J. (2005). Modeling patterns of probability calibration with Random Support Theory: Diagnosing case-based judgment. *Organizational Behavior and Human Decision Processes*, 97, 64–81.
- Brenner, L., Griffin, D. W., & Koehler, D. J. (2012). A case-based model of probability and pricing judgments: Biases in buying and selling uncertainty. *Management Science*. doi: 10.1287/mnsc.1110.1429.
- Brunswik, E. (1956). Perception and the representative design of psychological experiments. Berkeley: University of California Press.
- Camerer, C. F. (1987). Do biases in probability judgment matter in markets? Experimental evidence. *The American Economic Review*, 77, 981–998.
- Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. New York: Guilford.
- Carlson, B. W. (1990). Anchoring and adjustment in judgments under risk. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 665–676.
- Carroll, J. S. (1978). The effect of imagining an event on expectations for the event: An interpretation in terms of the availability heuristic. *Journal of Experimental Social Psychology*, 14, 88–96.
- Cervone, D., & Peake, P. (1986). Anchoring, efficacy, and action: The influence of judgmental heuristics on self-efficacy judgments and behavior. *Journal of Personality and Social Psychology*, 50, 492–501.
- Chaiken, S., & Trope, Y. (1999). *Dual-process theories in social psychology*. New York: Guilford Press.
- Chapman, G. B., & Johnson, E. J. (1999). Anchoring, activation and the construction of value. *Organizational Behavior and Human Decision Processes*, 79, 115–153.
- Cohen, J., & Hansel, C. E. M. (1956). *Risk and gambling*. New York: Philosophical Library.
- Cohen, J. (1960). *Chance, skill, and luck: The psychology of guessing and gambling*. Baltimore, MD: Penguin Books.
- Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment and uncertainty. *Cognition*, 58, 1–73.
- Critcher, C. R., & Gilovich, T. (2008). Incidental environmental anchors. *Journal of Behavioral Decision Making*, 21, 241–251.
- Dawes, R. M., Mirels, H. L., Gold, E., & Donahue, E. (1993). Equating inverse probabilities in implicit personality judgments. *Psychological Science*, 4, 396–400.
- De Finetti, B. (1970). Logical foundations and measurement of subjective probability. *Acta Psychologica*, 34, 129–145.
- Denes-Raj, V., & Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. *Journal of Personality and Social Psychology*, 66, 819–829.
- De Neys, W., & Goel, V. (2011). Heuristics and biases in the brain: Dual neural pathways for decision making. In O. Vartanian & D. R. Mandel (Eds.), *Neuroscience of decision making* (pp. 125–142). Hove, England: Psychology Press.
- Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17–52). New York: Wiley.
- Epley, N. (2004). A tale of tuned decks? Anchoring as accessibility and anchoring as adjustment. In D. J. Koehler & N. Harvey (Eds.), *The Blackwell handbook of judgment and decision making* (pp. 240–256). Oxford, England: Blackwell Publishers.
- Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Divergent processing of self-generated and experimenter-provided anchors. *Psychological Science*, 12, 391–396.
- Epley, N., & Gilovich, T. (2004). Are adjustments insufficient? *Personality and Social Psychology Bulletin*, 30, 447–460.
- Epley, N., & Gilovich, T. (2005). When effortful thinking influences judgmental anchoring: Differential effects of forewarning and incentives on self-generated and externally-provided anchors. *Journal of Behavioral Decision Making*, 18, 199–212.
- Epley, N., & Gilovich, T. (2006). The anchoring and adjustment heuristic: Why adjustments are insufficient. *Psychological Science*, 17, 311–318.
- Epstein, S. (1991). Cognitive-experiential self-theory: An integrative theory of personality. In R. Curtis (Ed.), *The self with others: Convergences in psychoanalytic, social, and personality psychology* (pp. 111–137). New York: Guilford Press.
- Evans, J. St. B.T. (2004). History of the dual process theory in reasoning. In K. I. Manktelow & M. C. Chung (Eds.), *Psychology of reasoning: Theoretical and historical perspectives* (pp. 241–266). Hove, England: Psychology Press.
- Evans, J. St. B.T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgment*. New York: Psychology Press.
- Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 59, 255–278.
- Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review*, 104, 301–318.
- Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, 90, 239–260.

- Fischhoff, B., Slovic, P., & Lichtenstein, S. (1978). Fault trees: Sensitivity of estimated failure probabilities to problem representations. *Journal of Experimental Psychology: Human Perception and Performance*, 3, 330–344.
- Fox, C. R. (1999). Strength of evidence, judged probability, and choice under uncertainty. *Cognitive Psychology*, 38, 167–189.
- Fox, C. R. (2006). The availability heuristic in the classroom: How soliciting more criticism can boost your course ratings. *Judgment and Decision Making*, 1, 86–90.
- Fox, C. R., & Hadar, L. (2006). Decisions from experience = sampling error + prospect theory: Reconsidering Hertwig, Barron, Weber & Erev (2004). *Judgment and Decision Making*, 1, 159–161.
- Gabrielcik, A., & Fazio, R. H. (1984). Priming and frequency estimation: A strict test of the availability heuristic. *Personality and Social Psychology Bulletin*, 10, 85–89.
- Gavanski, I., & Hui, C. (1992). Natural sample spaces and uncertain belief. *Journal of Personality and Social Psychology*, 63, 585–595.
- Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". *European Review of Social Psychology*, 2, 83–115.
- Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 129–161). New York: Wiley.
- Gigerenzer, G. (1998). Ecological intelligence: An adaptation for frequencies. In D. Dellarosa Cummins & C. Allen (Eds.), *The evolution of mind* (pp. 9–29). New York: Oxford University Press.
- Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 513–525.
- Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Erlbaum.
- Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life*. New York: The Free Press.
- Gilovich, T., & Griffin, D. W. (2010). Judgment and decision making. In D. T. Gilbert & S. T. Fiske (Eds.), *The handbook of social psychology* (pp. 542–588). New York: McGraw-Hill.
- Gilovich, T., & Savitsky, K. (2002). Like goes with like: The role of representativeness in erroneous and pseudo-scientific beliefs. In T. Gilovich, D. W. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 617–624). New York: Cambridge University Press.
- Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17, 295–314.
- Grether, D. (1992). Testing Bayes rule and the representativeness heuristic: Some experimental evidence. *Journal of Economics Behavior and Organization*, 17, 31–57.
- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantic. Volume 3: Speech acts* (pp. 225–242). New York: Seminar Press.
- Griffin, D. W., & Buehler, R. (1999). Frequency, probability, and prediction: Easy solutions to cognitive illusions? *Cognitive Psychology*, 38, 48–78.
- Griffin, D. W., Gonzalez, R., & Varey, C. A. (2001). The heuristics and biases approach to judgment under uncertainty. In A. Tesser & N. Schwarz (Eds.), *The Blackwell handbook of social psychology* (pp. 207–235). Oxford, England: Blackwell.
- Griffin, D. W., & Ross, L. (1991). Subjective construal, social inference, and human misunderstanding. In M. Zanna (Ed.), *Advances in experimental social psychology* (pp. 319–356). New York: Academic Press.
- Griffin, D. W., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24, 411–435.
- Helson, H. (1964). *Adaptation-level theory*. New York: Harper.
- Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, 15, 534–539.
- Hertwig, R., & Gigerenzer, G. (1999). The 'conjunction fallacy' revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, 12, 275–306.
- Hilton, D. J., & Slugoski, B. R. (2000). Judgment and decision making in social context: Discourse processes and rational inference. In T. Connolly, H. R. Arkes, & K. R. Hammond (Eds.), *Judgment and decision making: An interdisciplinary reader* (pp. 651–676). Cambridge, England: Cambridge University Press.
- Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology*, 3, 306–340.
- Jacoby, L. L., Woloshyn, V., & Kelley, C. (1989). Becoming famous without being recognized: Unconscious influences of memory produced by dividing attention. *Journal of Experimental Psychology: General*, 118, 115–125.
- Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, 21, 1161–1167.
- Jones, J. (2008, February 12). Truly functional foods. *PressRepublican.com*. Retrieved August 2011, from http://www.pressrepublican.com/0808_health/local_story_042224534.html
- Jones, S. K., Jones, K. T., & Frisch, D. (1995). Biases of probability assessment: A comparison of frequency and single-case judgments. *Organizational Behavior and Human Decision Processes*, 61, 109–122.
- Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus, Giroux.
- Kahneman, D., Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. W. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge, England: Cambridge University Press.
- Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In K. J. Holyoak, & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 267–293). New York: Cambridge University Press.
- Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, 93, 136–153.
- Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430–454.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237–251.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47, 263–291.
- Kahneman, D., & Tversky, A. (1982a). The psychology of preferences. *Scientific American*, 246, 160–173.

- Kahneman, D., & Tversky, A. (1982b). Variants of uncertainty. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 509–520). Cambridge, England: Cambridge University Press.
- Kahneman, D., & Tversky, A. (1983). Extension versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90, 293–315.
- Kahneman, D., & Varey, C. A. (1990a). Propensities and counterfactuals: The loser that almost won. *Journal of Personality and Social Psychology*, 59, 1101–1110.
- Kahneman, D., & Varey, C. A. (October, 1990b). *Mental contamination*. Paper presented at the Annual meeting of the Society for Experimental Social Psychology, Buffalo, NY.
- Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science*, 4, 533–550.
- Keynes, J. M. (1921). *A treatise on probability*. London: MacMillan.
- Koehler, J. J. (1996). The base-rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, 19, 1–53.
- Koehler, D. J., Brenner, L., & Griffin, D. W. (2002). The calibration of expert judgment: Heuristics and biases beyond the laboratory. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 686–715). New York: Cambridge University Press.
- Koffka, K. (1935). *Principles of Gestalt psychology*. New York: Harcourt, Brace & World.
- Krosnick, J. A., Li, F., & Lehman, D. R. (1990). Conversational conventions, order of information acquisition, and the effect of base rates and individuating information on social judgments. *Journal of Personality and Social Psychology*, 59, 1140–1152.
- Lopes, L. L. (1991). The rhetoric of irrationality. *Theory and Psychology*, 1, 65–82.
- Macchi, L. (1995). Pragmatic aspects of the base-rate fallacy. *Quarterly Journal of Experimental Psychology*, 48A, 188–207.
- Macleod, C., & Campbell, L. (1992). Accessibility and probability judgments: An experimental evaluation of the availability heuristic. *Journal of Personality and Social Psychology*, 63, 890–902.
- McGlone, M. S., & Tofighamkhsh, J. (2000). Birds of a feather flock conjointly(?): Rhyme as reason in aphorisms. *Psychological Science*, 11, 424–428.
- Meehl, P. E. (1954). *Clinical versus statistical prediction*. Minneapolis: University of Minnesota Press.
- Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, 52, 194–216.
- Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, 12, 269–275.
- Mischel, W. (1968). *Personality and assessment*. New York: Wiley.
- Miyamoto, J., Gonzalez, R., & Tu, S. (1995). Compositional anomalies in the semantics of evidence. *The Psychology of Learning and Motivation*, 32, 319–383.
- Mussweiler, T. (2002). The malleability of anchoring effects. *Experimental Psychology*, 49, 67–72.
- Mussweiler, T., & Strack, F. (1999). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology*, 35, 136–164.
- Mussweiler, T., & Strack, F. (2000). The use of category and exemplar knowledge in the solution of anchoring tasks. *Journal of Personality and Social Psychology*, 78, 1038–1052.
- Neisser, V. (1963). The multiplicity of thought. *British Journal of Psychology*, 54, 1–14.
- Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.
- Novemsky, N., Dhar, R., Schwarz, N., & Simonson, I. (2007). Preference fluency in choice. *Journal of Marketing Research*, 44, 347–356.
- Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences*, 12, 237–241.
- Oppenheimer, D. M., LeBoeuf, R. A., & Brewer, N. T. (2008). Anchors aweigh: A demonstration of cross-modality anchoring. *Cognition*, 206, 13–26.
- Oskamp, S. (1965). Overconfidence in case-study judgments. *Journal of Clinical and Consulting Psychology*, 29, 261–265.
- Plous, S. (1989). Thinking the unthinkable: The effect of anchoring on likelihood estimates of nuclear war. *Journal of Applied Social Psychology*, 19, 67–91.
- Pollack, I., & Decker, L. R. (1958). Confidence ratings, message reception, and the receiver operating characteristic. *Journal of the Acoustical Society of America*, 30, 286–292.
- Poulton, E. C. (1994). *Behavioral decision theory: A new approach*. Cambridge, England: Cambridge University Press.
- Ross, M., & Sicoly, F. (1979). Egocentric biases in availability and attribution. *Journal of Personality and Social Psychology*, 32, 880–892.
- Rothbart, M., Fulero, S., Jensen, C., Howard, J., & Birrell, P. (1978). From individual to group impressions: Availability heuristics in stereotype formation. *Journal of Experimental Social Psychology*, 14, 237–255.
- Rottenstreich, Y., & Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review*, 104, 406–415.
- Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.
- Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology*, 61, 195–202.
- Schwarz, N., Strack, F., Hilton, D., & Naderer, G. (1991). Base rates, representativeness, and the logic of conversation: The contextual relevance of “irrelevant” information. *Social Cognition*, 9, 67–84.
- Schwarz, N., & Vaughn, L. A. (2002). The availability heuristic revisited: Ease of recall and content of recall as distinct sources of information. In T. Gilovich, D. W. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 103–119). New York: Cambridge University Press.
- Sherman, S. J., Cialdini, R. B., Schwartzman, D. F., & Reynolds, K. D. (1985). Imagining can heighten or lower the perceived likelihood of contracting a disease: The mediating effect of ease of imagery. *Personality and Social Psychology Bulletin*, 11, 118–127.
- Simon, H. A. (1957). *Models of man: Social and rational*. New York: Wiley.
- Simmons, J. P., LeBoeuf, R. A., & Nelson, L. D. (2010). The effect of accuracy motivation on anchoring and adjustment: Do people adjust from provided anchors? *Journal of Personality and Social Psychology*, 99, 917–932.

- Skov, R. B., & Sherman, S. J. (1986). Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Social Psychology*, 22, 93–121.
- Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119, 3–22.
- Sloman, S. A., Over, D. E., Slovak, L., & Stiebel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes*, 91, 296–309.
- Sloman, S., Rottenstreich, Y., Wisniewski, E., Hadjichristidis, C., & Fox, C. (2004). Typical versus atypical unpacking and superadditive probability judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 573–582.
- Slovic, P., Finucane, M., Peters, E., & MacGregor, D. G. (2002). The affect heuristic. In T. Gilovich, D. W. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 397–420). Cambridge, England: Cambridge University Press.
- Slovic, P., Fischhoff, B., & Lichtenstein, S. (1982). Facts versus fears: Understanding perceived risk. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 463–489). New York: Cambridge University Press.
- Snyder, M., & Swann, W. B. (1978). Hypothesis-testing in social interaction. *Journal of Personality and Social Psychology*, 36, 1202–1212.
- Stanovich, K. E. (1999). *Who is rational? studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.
- Stanovich, K. E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 55–88). Oxford, England: Oxford University Press.
- Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, 8, 220–247.
- Strack, F., & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology*, 73, 437–446.
- Switzer, F., & Snizek, J. A. (1991). Judgment processes in motivation: Anchoring and adjustment effects on judgment and behavior. *Organizational Behavior and Human Decision Processes*, 49, 208–229.
- Tetlock, P. E. (2005). *Expert political judgment: How good is it? How can we know?* Princeton, NJ: Princeton University Press.
- Tversky, A., & Kahneman, K. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76, 105–110.
- Tversky, A., & Kahneman, K. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5, 207–232.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131.
- Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. (pp. 153–160). New York: Cambridge University Press.
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90, 293–315.
- Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, 101, 547–567.
- Wagenaar, W. A. (1972). Generation of random sequences by human subject: A critical survey of literature. *Psychological Bulletin*, 77, 65–72.
- Wason, P. C., & Evans, J. St. B. T. (1975). Dual processes in reasoning? *Cognition*, 3, 141–154.
- Whittlesea, B. W., & Leboe, J. P. (2000). The heuristic basis of remembering and classification: Fluency, generation, and resemblance. *Journal of Experimental Psychology: General*, 129, 84–106.
- Wong, K. F. E., & Kwong, J. Y. Y. (2000). Is 7300 m equal to 7.3 km? Same semantics but different anchoring effects. *Organizational Behavior and Human Decision Processes*, 82, 314–333.
- Wright, W. F., & Anderson, U. (1989). Effects of situation familiarity and financial incentives on use of the anchoring and adjustment heuristic probability assessment. *Organizational Behavior and Human Decision Processes*, 44, 68–82.
- Yates, J. F., & Estin, P. A. (November, 1996). *Training good judgment*. Paper presented at the Annual Meeting of the Society for Judgment and Decision Making, Chicago.
- Yates, J. F., Lee, J., Shinotsuka, H., Patalano, A. L., & Sieck, W. R. (1998). Cross-cultural variations in probability judgment accuracy: Beyond general knowledge overconfidence? *Organizational Behavior and Human Decision Processes*, 74, 89–117.
- Zukier, H., & Pepitone, A. (1984). Social roles and strategies in prediction: Some determinants of the use of base rate information. *Journal of Personality and Social Psychology*, 47, 349–360.

Cognitive Hierarchies and Emotions in Behavioral Game Theory

Colin F. Camerer and Alec Smith

Abstract

Until recently, game theory was not focused on cognitively plausible models of choices in human strategic interactions. This chapter describes two new approaches that do so. The first approach, cognitive hierarchy modeling, assumes that players have different levels of partially accurate representations of what others are likely to do, which vary from heuristic and naïve to highly sophisticated and accurate. There is reasonable evidence that this approach explains choices (better than traditional equilibrium analysis) in dozens of experimental games and some naturally occurring games (e.g., a Swedish lottery, auctions, and consumer reactions to undisclosed quality information about movies). Measurement of eye tracking and functional magnetic resonance imaging (fMRI) activity during games is also suggestive of a cognitive hierarchy. The second approach, psychological games, allows value to depend upon choice consequences *and* on beliefs about what will happen. This modeling framework can link cognition and emotion, and express social emotions such as “guilt.” In a psychological game, guilt is modeled as the negative emotion of knowing that another person is unpleasantly surprised that your choice did not benefit her (as she had expected). Our hope is that these new developments in a traditionally cognitive field (game theory) will engage interest of psychologists and others interested in thinking and social cognition.

Key Words: bounded rationality, cognitive hierarchy, emotions, game theory, psychological games, strategic neuroscience

Introduction

This chapter is about cognitive processes in strategic thinking. The theory of games provides the most comprehensive framework for thinking about the valued outcomes that result from strategic interactions. The theory specifies how “players” (that’s game theory jargon) might choose high-value strategies to guess likely choices of other players. Traditionally, game theory has been focused on finding “solutions” to games based on highly mathematical conceptions of rational forecasting and choice. More recently (starting with Camerer, 1990), behavioral game theory models have extended the rational theories to include stochastic response; limits on inferring

correctly what other players will do; social emotions and considerations such as guilt, anger, reciprocity, or social image; and modulating factors, including inferences about others’ intentions. Two general behavioral models that might interest cognitive psychologists are the focus of this chapter¹: cognitive hierarchy modeling and psychological game theory.

Conventional game theory is typically abstract, mathematically intimidating, computationally implausible, and algorithmically incomplete. It is therefore not surprising that conventional tools have not gained traction in cognitive psychology. Our hope is that the more psychologically plausible behavioral variants could interest cognitive

psychologists. Once limited strategic thinking is the focus, questions of cognitive representation, categorization of different strategic structures, and the nature of social cognition and how cooperation is achieved all become more interesting researchable questions. The question of whether people are using the decision-making algorithms proposed by these behavioral models can also be addressed with observables (such as response times and eye tracking of visual attention) familiar in cognitive psychology. Numerical measures of value and belief derived in these theories can also be used as parametric regressors to identify candidate brain circuits that appear to encode those measures. This general approach has been quite successful in studying simpler nonstrategic choice decisions (Glimcher, Camerer, Fehr, & Poldrack, 2008) but has been applied infrequently to games (see Bhatt & Camerer, 2011).

What Is a Game?

Game theory is the mathematical analysis of strategic interaction. It has become a standard tool in economics and theoretical biology, and it is increasingly used in political science, sociology, and computer science. A game is mathematically defined as a set of players, descriptions of their information, a fixed order of the sequence of choices by different players, and a function mapping players' choices and information to outcomes. Outcomes may include tangibles like corporate profits or poker winnings, as well as intangibles like political gain, status, or reproductive opportunities (in biological and evolutionary psychology models). The specification of a game is completed by a payoff function that attaches a numerical value or "utility" to each outcome.

The standard approach to the analysis of games is to compute an equilibrium point, a set of strategies for each player which are simultaneously best responses to one another. This approach is due originally to John Nash (1950), building on earlier work by Von Neumann and Morgenstern (1944). Solving for equilibrium mathematically requires solving simultaneous equations in which each player's strategy is an input to the other player's calculation of expected payoff. The solution is a collection of strategies, one for each player, where each player's strategy maximizes his expected payoff given the strategies of the other players.

From the beginning of game theory, how equilibrium might arise has been the subject of ongoing discussion. Nash himself suggested that equilibrium beliefs might resolve from changes in "mass action"

as populations learn about what others do and adjust their strategies toward optimization.²

More recently game theorists have considered the epistemic requirements for Nash equilibrium by treating games as interactive decision problems (cf. Brandenburger, 1992). It turns out that Nash equilibrium for n -player games requires very strong assumptions about the players' *mutual* knowledge: that all players share a common prior belief about chance events, know that all players are rational, and know that their beliefs are common knowledge (Aumann & Brandenburger, 1995).³ The latter requirement implies that rational players be able to compute beliefs about the strategies of coplayers and all states of the world, beliefs about beliefs, and so on, ad infinitum.

Two Behavioral Approaches: Cognitive Hierarchy and Psychological Games

Cognitive hierarchy (CH) and psychological games (PG) models both modify assumptions from game theory to capture behavior more realistically.

The CH approach assumes that boundedly rational players are limited in the number of interpersonal iterations of strategic reasoning they can (or choose) to do. There are five elements to any CH predictive model:

1. A distribution of the frequency of level types $f(k)$
2. Actions of level-0 players
3. Beliefs of level- k players (for $k = 1, 2, \dots$) about other players
4. Assessing expected payoffs based on beliefs in (3)
5. A stochastic choice response function based on the expected payoffs in (4)

The typical approach is to make precise assumptions about elements (1–5) and see how well that specific model fits experimental data from different games. Just as in testing a cooking recipe, if the model fails badly then it can be extended and improved.

In Camerer, Ho, and Chong (2004), the distribution of level- k types is assumed to follow a Poisson distribution with a mean value τ . Once the value of τ is chosen, the complete distribution is known. The Poisson distribution has the sensible property that the frequencies of very high-level types k drop off quickly for higher values of k . (For example, if the average number of thinking steps $\tau = 1.5$, then less than 2% of players are expected to do five or more steps of thinking.)

To further specify the model, level-0 types are usually assumed to choose each strategy equally often.⁴ In the CH approach, level- k players know the correct proportions of lower-level players, but they do not realize there are other even higher-level players (perhaps reflecting overconfidence in relative ability). An alternative assumption (called “level- k ” modeling) is that a level- k player thinks *all* other players are at level $k - 1$.

Under these assumptions, each level of player in a hierarchy can then compute the expected payoffs to different strategies: Level 1’s compute their expected payoff (knowing what level 0’s will do); level 2’s compute the expected payoff given their guess about what level 1’s and 0’s do, and how frequent those level types are; and so forth. In the simplest form of the model, players choose the strategy with the highest expected payoff (the “best response”); but it is also easy to use a logistic or power stochastic “better response” function (e.g., Luce, 1959). Because the theory is hierarchical, it is easy to program and solve numerically using a “loop.”

Psychological games models assume that players are rational in the sense that they maximize their expected utility given beliefs and the utility functions of the other players. However, in psychological games models, payoffs are allowed to depend directly upon players’ beliefs, their beliefs about their coplayers’ beliefs, and so on, a dependence that is ruled out in standard game theory. The incorporation of belief-dependent motivations makes it possible to capture concerns about intentions, social image, or even emotions in a game-theoretic framework. For example, in psychological games one person, Conor(C) might be delighted to be surprised by the action of another player, Lexie (L). This is modeled mathematically as C liking when L’s strategy is different than what he (C) expected L to do. Some of these motivations are naturally construed as social emotions, such as guilt (e.g., a person feels bad choosing a strategy which harmed another person P who did not expect it, and she feels less bad if P did expect it).

Of the two approaches, CH and level- k modeling are easy to use and apply to empirical settings. Psychological games are more general, applying to a broader class of games, but are more difficult to adapt to empirical work.

The CH Model

The next section gives some motivating empirical examples of the wide scope of games to which the theory has been applied with some success

(including two kinds of field data), and consistency with data on visual fixation and fMRI. The CH approach is appealing as a potential cognitive algorithm for four reasons:

1. It appears to fit a lot of experimental data from many different games better than equilibrium predictions do (e.g., Camerer et al., 2004; Crawford, Costa-Gomes, & Iribarri, 2010).
2. The specification of how thinking works and creates choices invites measurement of the thinking process with response times, visual fixations on certain payoffs, and transitions between particular payoffs.
3. The CH approach introduces a concept of skill into behavioral game theory. In the CH model, the players with the highest thinking levels (higher k) and most responsive choices (higher λ) are implicitly more skilled. (In equilibrium models, all players are perfectly and equally skilled.)

Next we will describe several empirical games that illustrate how CH reasoning works.

Example 1: p-beauty contest

A simple game that illustrates apparent CH thinking has come to be called the “p-beauty contest game” (or PBC). The name comes from a famous passage in John Maynard Keynes’s book *The General Theory of Employment, Interest and Money*. Keynes wrote:

Professional investment may be likened to those newspaper competitions in which the competitors have to pick out the six prettiest faces from a hundred photographs, the prize being awarded to the competitor whose choice most nearly corresponds to the average preferences of the competitors as a whole; so that each competitor has to pick, not those faces which he himself finds prettiest, but those which he thinks likeliest to catch the fancy of the other competitors, all of whom are looking at the problem from the same point of view. It is not a case of choosing those which, to the best of one’s judgment, are really the prettiest, nor even those which average opinion genuinely thinks the prettiest. We have reached the third degree where we devote our intelligences to anticipating what average opinion expects the average opinion to be. And there are some, I believe, who practice the fourth, fifth and higher degrees. (p. 156)

In the experimental PBC game, people choose numbers from 0 to 100 simultaneously without talking.⁵ The person whose number is closest to p times the average wins a fixed price.

A typical interesting value of p is $2/3$. Then the winner wants to be two-thirds of the way between the average and zero. But, of course, the players all know the other players want to pick $2/3$ of the average. In a Nash equilibrium, everyone accurately forecasts that the average will be X and also chooses a number which is $(2/3)X$. This implies $X = (2/3)X$ or $X^* = 0$.

Intuitively, suppose you had no idea what other people would do, so you chose $2/3$ of $50 = 33$. This is a reasonable choice but is not an equilibrium, since choosing 33 while anticipating 50 leaves a gap between expected behavior of others and likely behavior by oneself. So a person who thinks, “Hey! I’ll pick 33” should then think (to adhere to the equilibrium math), “Hey! *They’ll* pick 33” and then pick 22. This process of imagining, choosing, and revising does not stop until everyone expects 0 to be chosen and also picks 0.

Figure 18.1 shows some data from this game played with experimental subjects and in newspaper and magazine contests (where large groups play for a single large prize). There is some evidence of “spikes” in numbers corresponding to $50p$, $50p^2$ and so on.

Example 2: Betting on selfish rationality of others

Another simple illustration of the CH theory is shown in Table 18.1. In this game a row and column player choose from one of two strategies, T or B (for row) or L or R (for column). The column player always gets 20 for choosing L and 18 for choosing R. The row player gets either 30 or 10 from T, and a sure 20 from B.

If the column player is trying to get the largest payoff, she should always choose L (it guarantees 20 instead of 18). The strategy L is called a “strictly dominant strategy” because it has the highest payoff for every possible choice by the row player.

The row player’s choice is a little trickier. She can get 20 for sure by choosing B. Choosing T is taking a social gamble. If she is confident the column player will try to get 20 and choose L, she should *infer* that $P(L)$ is high. Then the expected value of T is high and she should choose T. However, this inference is essentially a bet on the selfish rationality of the other player. The row player might think the column player will make a mistake or is spiteful (and prefers the (10,18) cell because she gets less absolute payoff but a higher relative payoff compared to the row player). There is a crucial cognitive difference in playing L—which is the right strategy if you want the most money—and playing T—which is the right strategy if you are willing to bet that other players are very likely to choose L because they want to earn the most money.

What does the CH approach predict here? Suppose level-0 players randomize between the two strategies. If $\tau = 1.5$, then $f(0|\tau = 1.5) = .22$. Then half of the level-0 players will choose column R and row B, which is .11% of the whole group.

Level-1 players always choose weakly dominant strategies, so they pick column L (in fact, all higher level column players do, too). Since level-1 row players think L and R choices are equally likely, their expected payoff from T is $30(.5) + 10(.5) = 20$, which is the same as the B payoff; so we assume they randomize equally between T and B. Since $f(1|\tau = 1.5) = .33$, this means the unconditional total frequency of B play for the first two levels is $.11 + .33/2 = .27$.

Level-2 row players think the relative proportions of lower types are $g_2(0) = .22/(.22 + .33) = .40$ and

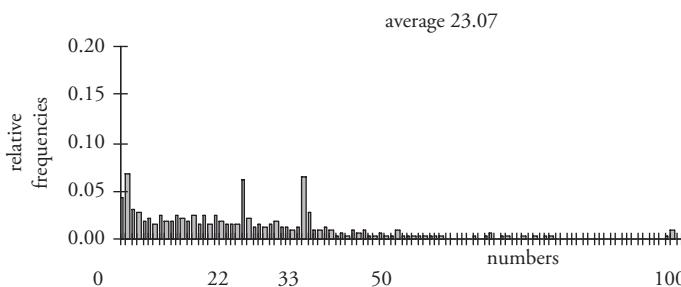


Fig. 18.1 Choices in “ $2/3$ of the average” game (Bosch-Domenech, Garcia-Montalvo, Nagel, & Satorra, 2002).

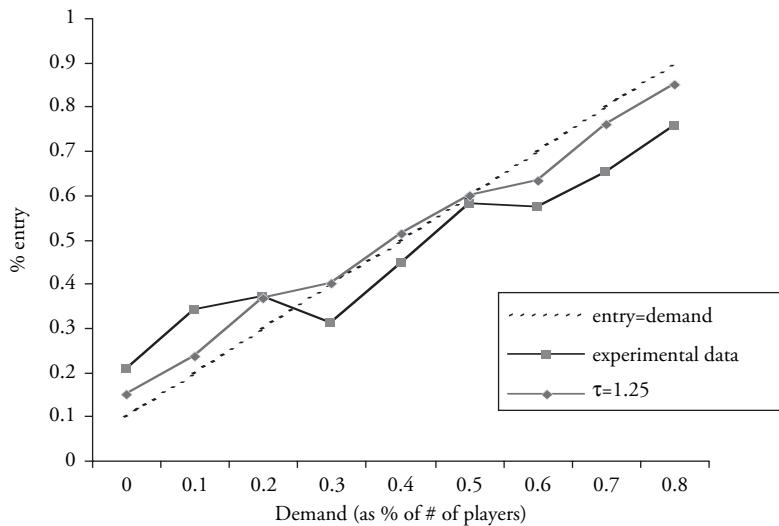


Fig. 18.2 Predicted and observed behavior in entry games (Camerer, Ho, & Chong, 2004). See color figure.

$g_2(1) = .33/(.22 + .33) = .60$. They also think the level 0's play either L or R, but the level 1's choose L for sure. Together, this implies that they believe there is a .20 chance the other person will choose R ($= .5(.40) + 0(.60)$) and an .80 chance they will choose L. With these odds, they prefer to choose T. That is, they are sufficiently confident the other player will “figure it out” and choose the self-serving L that T becomes a good bet to yield the higher payoff of 30.

Putting together all the frequencies $f(k)$ and choice percentages, the overall expected proportion of column R play is .11 and row B play is .27. Note that these proportions go in the direction of the Nash prediction (which is zero for both), but they account more precisely for the chance of mistakes

and misperceptions. Importantly, choices of R should be less common than choices of B. R choices are just careless, while B choices might be careless *or* might be sensible responses to thinking there are a lot of careless players.

Table 18.1 shows that some (unpublished) data from Caltech undergraduate classroom games (for money) over 3 years are generally close to the CH prediction. The R and B choice frequencies are small (as both Nash and CH predict), but B is more common than R.

One potential advantage of CH modeling is that the same general process could apply to games with different economic structures. In both of the two examples, a Nash equilibrium choice can be derived

Table 18.1. Payoffs in Betting Game, Predictions (Nash and CH), and Results From Classroom Demonstrations in 2006–2008

	L	R	Predictions		Data	
			Nash	CH	2006 + 07 + 08	Average
T	30, 20	10, 18	1.00	.73	.81 + .86 + .78	.82
B	20, 20	20, 18	.00	.27	.19 + .14 + .22	.18
Nash	1.00	0				
CH	.89	.11				
2006 + 07 + 08	.95 + .95 + .75	.05 + .05 + .25				
Average	.88	.12				

Note. Upper left is the unique Nash equilibrium.

Table 18.2. Payoffs From Hand T Choice in a “Matching Pennies” Game, Predictions, and Data

		Predictions						
		H	T	Nash	CH	Levels 1–2	Levels 3–4	Data
H		2,0	0,1	.50	.68	1	0	.72
T		0,1	1,0	.50	.32	0	1	.28
Nash		.33	.67					
CH		.26	.74					
Data		.33	.67					

by repeated application of the principle of eliminating “weakly dominated” strategies (i.e., strategies that are never better than another dominating strategy, for all choices by other people, and is actually worse for some choices by others). Hence, these are called “dominance solvable” games. Indeed, the beauty-contest example is among those that motivated CH modeling in the first place, since each step of reasoning corresponds to one more step in deletion of dominated strategies.

Table 18.2 shows predictions and results for an entirely different type of game, called “asymmetric matching pennies.” In this game the row player earns points if the choices match (H,H) or (T,T). The column player wins if they mismatch. There is no pair of strategies that are best responses to each other, so the equilibrium requires choosing a probabilistic “mixture” of strategies. Here, equilibrium analysis makes a bizarre prediction: The row player should choose H and T equally often, while the column player should shy away from H (as if preventing Row from getting the bigger payoff of 2) and choose T 2/3 of the time. (Even more strangely: If the 2 payoff is $x > 1$ in general, then the mixture is always 50–50 for the row player, and is $x/(x+1)$ on T for the column player! That is, in theory changing the payoff of 2 *only* affects the column player, and it does not affect the row player who might earn that payoff.)

The CH approach works differently.⁶ The lower level row players (1–2) are attracted to the possible payoff of 2, and choose H. However, the low-level column players switch to T, and higher level row players (levels 3–4) figure this out and switch to T. The predicted mixture (for $\tau = 1.5$) is actually rather close to the Nash prediction for the column player ($P(T) = .74$ compared to Nash .67), since the

higher-level types choose T more and not H. And indeed, data from column player choices in experiments are close to both predictions. The CH mixture of row play, averaged across type frequencies, is $P(H) = .68$, close to the data average of .72. Thus, the reasonable part of the Nash prediction, which is lopsided play of T and H by column players, is reproduced by CH and is consistent with the data. The unreasonable part of the Nash prediction, that row players choose H and T equally often, is not reproduced and the differing CH prediction is more empirically accurate.

Entry Games

In simple “entry” games, N players simultaneously choose whether to enter a market with demand C. If they stay out, they earn a fixed payoff (\$.50). If they enter, then all the entrants earn \$1 if there are C or fewer entrants, and they earn 0 if there are more than C entrants. It is easy to see that the equilibrium pattern of play is for *exactly* C people to enter; then they each earn \$1 and those who stay out earn \$.50. If one of the stayer-outers switched and entered, she would tip the market and cause the C + 1 entrants to earn 0. Since this would lower her own payoff, she will stay put. So the pattern is an equilibrium.

However, there is a problem remaining (it’s a common one in game theory): How does the group collectively decide, without talking, *which* of the C people enter and earn \$1? Everybody would like to be in the select group of C entrants if they can; but if too many enter they all suffer.⁷ This is a familiar problem of “coordinating” to reach one of many different equilibria.

The first experiments on this type of entry game were done by a team of economists (James Brander and Richard Thaler) and a psychologist, Daniel Kahneman. They were never fully published but were described in a chapter by Kahneman (1988). Kahneman says they were amazed how close the number of total entrants was to the announced demand C (which varied over trials). “To a psychologist,” he wrote, “it looked like magic.” Since then, a couple of dozen studies have explored variants of these games and reported similar degrees of coordination (e.g., Duffy & Hopkins, 2005).

Let’s see if cognitive hierarchy can produce the magic. Suppose level-0 players enter and stay out equally often, and ignore C. If level-1 players anticipate this, they will think there are too many entrants for $C < (N/2)$ and too few if $C > (N/2) - 1$. Level-1 players will therefore enter at high values

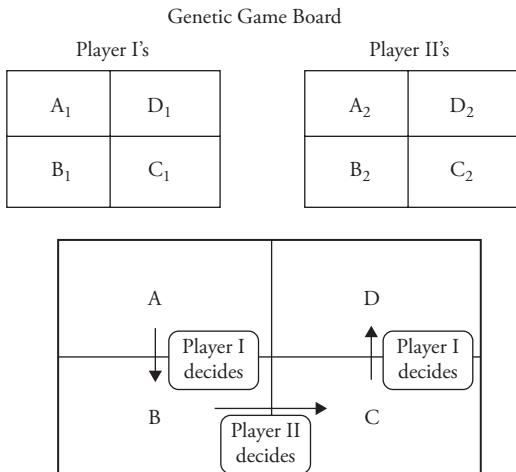


Fig. 18.3 The game board from Hedden and Zhang (2002). Generic game structure with arrows indicating direction of play.

of C. Notice that level-1 players are helping the group move toward the equilibrium. Level 1's undo the damage done by the level 0's, whoever enter at low C, by staying out, which reduces the overall entry rate for low C. They also exploit the opportunity that remains for high C, by entering, which increases the overall entry rate. Combining the two levels, there will be less entry at low C and more entry at high C (it will look like a step function; see Camerer et al., 2004).

Furthermore, it turns out that adding higher-level thinkers continues to push the population profile toward an overall entry level that is close to C. The theory makes three sharp predictions: (1) Plotting entry rates (as a % of N) against C/N should yield a regressive line which crosses at (.5, .5). (2) Entry rates should be too high for C/N < .5 and too low for C/N > .5. (3) Entry should be increasing in C, and relatively close, even without any learning at all (e.g., in the first period of the game).

Figure 18.2 illustrates a CH model prediction with $\tau = 1.25$, single-period data with no feedback from Camerer et al. (2004), and the equilibrium (a 45-degree line). Except for some nonmonotonic dips in the experimental data (easily accounted for by sampling error), the predictions are roughly accurate.

The point of this example is that approximate equilibration can be produced, as if by “magic,” purely from cognitive hierarchy thinking without any learning or communication needed. These data are not solid proof that cognitive hierarchy reasoning is occurring in this game, but it does show how, in principle, the cognitive hierarchy approach can

explain both deviations from Nash equilibrium (in the beauty contest, betting, and matching pennies games that were described earlier), and also surprising conformity to Nash equilibrium (in this entry game).

Private Information

The trickiest class of games we will discuss, briefly, involve “private information.” The standard modeling approach is to assume there is a hidden variable, X , which has a possible distribution $p(X)$ that is commonly known to both players (Harsanyi, 1967).⁸ The informed player I knows the exact value X from the distribution, and both players know that only I knows the value. For example, in card games like poker, players know the possible set of cards their opponent might have, and they know that the opponent knows exactly what the cards are.

The cognitive challenge that is special to private information games is to infer what a player’s actions, whether they are actually taken or hypothetical, might reveal about his information. Various experimental and field data indicate that some players are not very good at inferring hidden information from observed action (or anticipating the inferable information).

A simple and powerful example is the “acquire-a-company” problem introduced in economics by Akerlof (1970) and studied empirically by Bazerman and Samuelson (1983). In this game, a privately held company has a value which is perceived by outsiders to be uniformly distributed from 0 to 100 (i.e., all values in that range are equally likely). The company knows its exact value, and outsiders know that the company knows (due to the common prior assumption).

A bidder can operate the company much better, so that whatever the hidden value V is, it is worth $1.5V$ to them. They make a take-it-or-leave-it (Boulwarean) offer of a price P . The bargaining could hardly be simpler: The company sells if the price P is above “hidden” value V —which the bidder knows that the company knows—and keeps the company otherwise. The bidder wants to maximize the expected “surplus” gain between the average of the values $1.5V$ he is likely to receive and the price.

What would you bid? The optimal bid is surprising, though the algebra behind the answer is not too hard. The chance of getting the company is the chance that V is less than P , which is $P/100$ (e.g., if $P = 60$ then 60% of the time the value is below P and the company changes hands). If the

company is sold, then the value must be below P , so the expected value to the seller is the average of the values in the interval $[0, P]$, which is $P/2$. The net expected value is therefore $(P/100)$ times expected profit *if sold*, which is $1.5*(P/2) - P = -1/4P$. There is no way to make a profit on average. The optimal bid is zero!

However, typical distributions of bids are between 50 and 75. This results in a “winner’s curse” in which bidders “win” the company but fail to account for the fact that they only won because the company had a low value. This phenomenon was first observed in field studies of oil-lease bidding (Capen et al., 1971) and has been shown in many lab and field data sets since then. The general principle that people have a hard time guessing the implications of private information for actions others will take shows up in many economic settings (a kind of strategic naivete; e.g., Brocas et al., 2009).

The CH approach can easily explain strategic naivete as a consequence of level-1 behavior. If level-1 players think that level-0 players’ choices do not depend on private information, then they will ignore the link between choices and information.

Eye-Tracking Evidence

A potential advantage of cognitive hierarchy approaches is that cognitive measures associated with the algorithmic steps players are assumed to use, in theory, could be collected along with choices. For psychologists this is obvious, but, amazingly, it is a rather radical position in economics and most areas of game theory!

The easiest and cheapest method is to record what information people are looking at as they play games. Eye-tracking measures visual fixations using video-based eye tracking, typically every 5–50 msec. Cameras look into the eye and adjust for head motion to guess where the eyes are looking (usually with excellent precision). Most eye trackers record pupil dilation as well, which is useful as a measure of cognitive difficulty or arousal.

Since game theory is about interactions among two or more people, it is especially useful to have a recording technology that scales up to enable recording of several people at the same time. One widely used method is called “Mouselab.” In Mouselab, information that is used in strategic computations, in theory, is hidden in labeled boxes, which “open up” when a mouse is moved into them.⁹

Several studies have shown that lookup patterns often correspond roughly, and sometimes quite closely,

to different numbers of steps of thinking. We’ll present one example (see also Crawford et al., 2010).

Example 1: Alternating-offer bargaining

A popular approach to modeling bargaining is to assume that players bargain over a known sum of joint gain (sometimes called “surplus,” like the valuable gap between the highest price a buyer will pay and the lowest price a seller will accept). However, as time passes, the amount of joint gain “shrinks” due to impatience or other costs. Players alternate, making offers back and forth (Rubinstein, 1982).

A three-period version of this game has been studied in many experiments. The amount divided in the first round is \$5, which then shrinks to \$2.50, \$1.25, and 0 in later rounds (the last round is an “ultimatum game”). If players are selfish and maximize their own payoffs, and believe that others are too, the “subgame perfect” equilibrium (SPE) offer by the first person who offers (player 1), to player 2, should be \$1.25. However, deriving this offer either requires some process of learning or communication, or an analysis using “backward induction” to deduce what offers would be made and accepted in all future rounds, then working back to the first round. Early experiments showed conflicting results in this game. Neelin et al. (1988) found that average offers were around \$2, and many were equal splits of \$2.50 each. Earlier, Binmore et al. (1985) found similar results in the first round of choices but also found that a small amount of experience with “role reversal” (player 2’s switching to the player 1 first-offer position) moved offers sharply toward the SPE offer of \$1.25. Other evidence from simpler ultimatum games showed that people seem to care about fairness and are willing to reject a \$2 offer out of \$10 about half the time, to punish a bargaining partner they think has been unfair and greedy (Camerer, 2003).

So are the offers around \$2 due to correct anticipation of fairness-influenced behavior or to limited understanding of how the future rounds of bargaining might shape reactions in the first round? To find out, Camerer et al. (1993) and Johnson et al. (2002) did the same type of experiment but hid the amounts being bargained over in each round in boxes that could be opened, or “looked up,” by moving a mouse into those boxes (an impoverished experimenter’s version of video-based eye tracking).

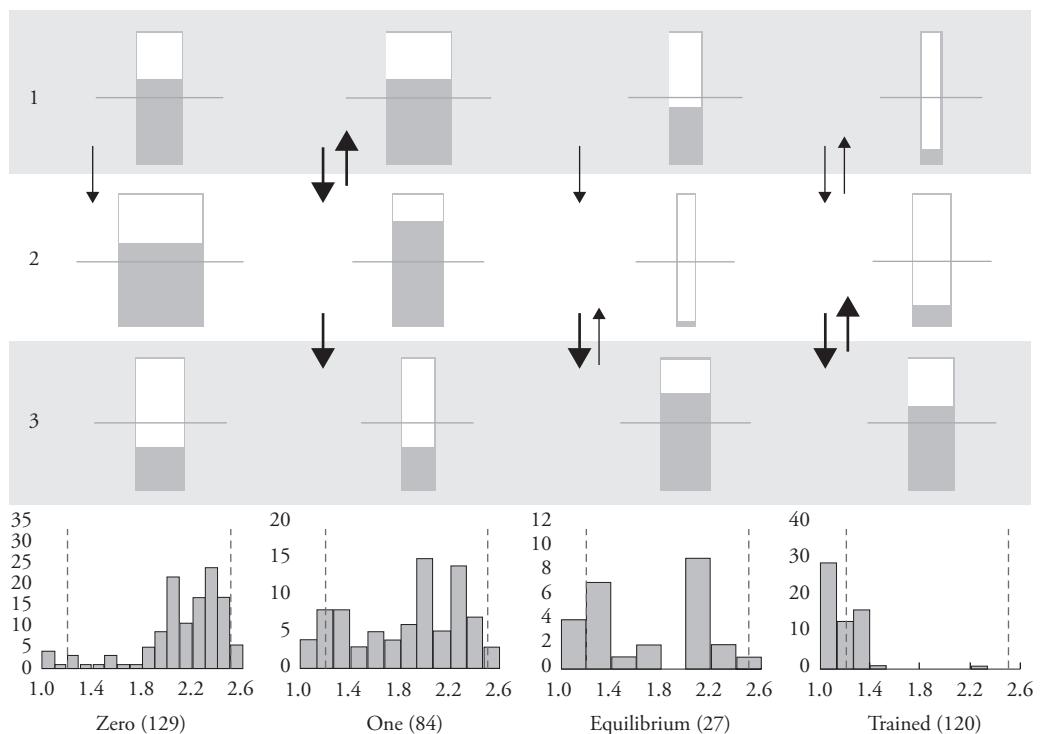


Fig. 18.4 An icon graph of visual attention in three rounds of bargaining (1, 2, and 3) and corresponding distributions of offers. Each column represents a different “type” of person-trial classified by visual attention. See color figure.

They found that most people who offered amounts between the equal split of \$2.50 and the SPE of \$1.25 were not looking ahead at possible future payoffs as backward induction requires. In fact, in 10%–20% of the trials the round 2 and round 3 boxes were not opened at all!

Figure 18.4 illustrates the basic results. The top rectangular “icon graphs” visually represent the relative amounts of time bargainers spent looking at each payoff box (the shaded area) and numbers of different lookups (rectangle width). The bold arrows indicate the relative number of transitions from one box to the next (with averages of less than one transition omitted).

Each column represents a group of trials that are preclassified by lookup patterns. The first column ($N = 129$ trials) averages people who looked more often at the period 1 box than at the future period boxes (indicating “level-0” planning). The second column ($N = 84$) indicates people who looked longer at the second box than the first and third (indicating “level One” planning with substantial focus one step ahead). The third column ($N = 27$) indicates the smaller number of “equilibrium” trials

in which the third box is looked at the most. Note that in the level-1 and Equilibrium trials, there are also many transitions between boxes one and two, and boxes two and three, respectively. Finally, the fourth and last column shows people who were briefly trained in backward induction, then played a computerized opponent that (they were told) planned ahead, acted selfishly, and expected the same from its opponents.

The main pattern to notice is that offer distributions (shown at the bottom of each column) shift from right (fairer, indicated by the right dotted line) to left (closer to selfish SPE, the left dotted line) as players literally look ahead more. The link between lookups and higher-than-predicted offers clearly shows that offers above the SPE, in the direction of equal splits of the first round amount, are partly due to limits on attention and computation about future values. Even in the few equilibrium trials, offers are bimodal, clustered around \$1.25 and \$2.00. However, offers are rather tightly clustered around the SPE prediction of \$1.25 in the “trained” condition. This result indicates, importantly, that backward induction is not actually that cognitively challenging to execute (after instruction, they can

easily do it), but instead it is an unnatural heuristic that does not readily spring to the minds of even analytical college students.

fMRI Evidence

Several neural studies have explored which brain regions are most active in different types of strategic thinking. The earliest studies showed differential activation when playing a game against a computer compared to a randomized opponent (e.g., Coricelli & Nagel, 2009; Gallagher et al., 2002; McCabe et al., 2001).

One of the cleanest results, and an exemplar of the composite picture emerging from other studies, is from Coricelli and Nagel's (2009) study of the beauty contest game. Their subjects played 13 different games with different target multipliers p (e.g., $p = 2/3, 1, 3/2$, etc.). On each trial, subjects chose numbers in the interval [0,100] playing against either human subjects or against a random computer opponent. Using behavioral choices, most subjects can be classified into either level 1 ($n = 10$; choosing p times 50) or level 2 ($n = 7$; choosing p times p times 50, as if anticipating the play of level-1 opponents).

Figure 18.7 shows brain areas that were differentially active when playing human opponents compared to computer opponents, and in which that human-computer differential is larger in level-2 players compared to level-1 players. The crucial areas are bilateral temporoparietal junction (TPJ), MPFC/paracingulate, and VMPFC.¹⁰ These regions are thought to be part of a general mentalizing circuit, along with posterior cingulate regions (Amodio & Frith, 2006).

In recent studies, at least four areas are reliably activated in higher-level strategic thinking: dorsomedial prefrontal cortex (DMPFC), precuneus/posterior cingulate, insula, and dorsolateral prefrontal cortex (DLPFC). Next we summarize some of the simplest results.

DMPFC activity is evident in Figure 18.7. It is also active in response to nonequilibrium choices (where subjects' guesses about what others will do are wrong; Bhatt & Camerer, 2005) and uncertainty about strategic sophistication of an opponent (Yoshida et al., 2009). In addition, DMPFC activity is related to the "influence value" of current choices on future rewards, filtered through the effect of a person's future choices on an opponent's future choices (Hampton, Bossaerts, & O'Doherty, 2008; aka "strategic teaching"; Camerer, Ho, & Chong, 2002; Chong, Camerer, & Ho, 2006). Amodio and

Frith (2006) suggest an intriguing hypothesis: that mentalizing-value activation for simpler to more complex action value computations is differentially located along a posterior-to-anterior (back-to-front) gradient in DMPFC. Indeed, the latter three studies show activation roughly in a posterior-to-anterior gradient (Tailarach $y = 36, 48, 63$; and $y = 48$ in Coricelli & Nagel, 2009) that corresponds to increasing complexity.

Activity in the precuneus (adjacent to posterior cingulate) is associated with economic performance in games ("strategic IQ;" Bhatt & Camerer, 2005) and difficulty of strategic calculations (Kuo et al., 2009). Precuneus is a busy region, with reciprocal connections to MPFC, cingulate, and DLPFC (Cavanna & Trimble, 2006). It is also activated by a wide variety of higher-order cognitions, including perspective-taking and attentional control (as well as the "default network" active at rest; see Bhatt & Camerer, 2011). It is likely that precuneus is not activated in strategic thinking, per se, but only in special types of thinking that require taking unusual perspectives (e.g., thinking about what other people will do) and shifting mental attention back and forth.

The insula is known to be involved in interoceptive integration of bodily signals and cognition. Disgust, physical pain, empathy for others in pain, and pain from social rejection activate insula (Eisenberger et al., 2003; Kross et al., 2011). Financial uncertainty (Preuschoff, Quartz, & Bossaerts, 2008), interpersonal unfairness (Hsu et al., 2008; Sanfey et al., 2003), avoidance of guilt in trust games (Chang et al., 2011), and "coaxing" or second-try signals in trust games also activate insula. In strategic studies, Bhatt and Camerer (2005) found that higher insula activity is associated with lower strategic IQ (performance).

The DLPFC is involved in working memory, goal maintenance, and inhibition of automatic prepotent responses. Differential activity there is also associated with the level of strategic thinking (Yoshida et al., 2009) with stronger response to human opponents in higher-level strategic thinkers (Coricelli & Nagel, 2009), and with maintaining level-2 deception in bargaining games (Bhatt et al., 2010)¹¹

Do Thinking Steps Vary With People or Games?

To what extent do steps of thinking vary systematically across people or game structures? From a cognitive point of view, it is likely that there is *some* intrapersonal stability because of differences in working memory, strategic savvy, exposure to game

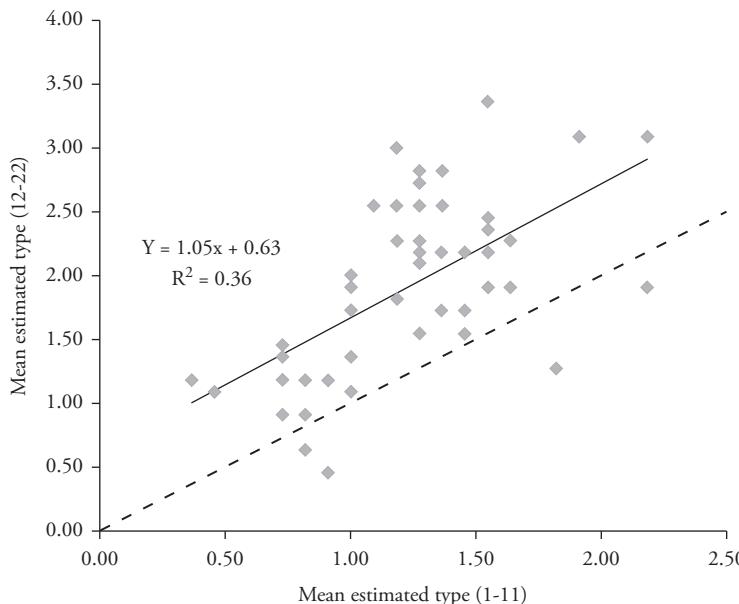


Fig. 18.5 Estimated strategic level types for each individual in two sets of 11 different games (Chong, Camerer, and Ho, 2005). Estimated types are correlated in two sets ($r = .61$).

theory, experience in sports betting or poker, task motivation, and so on. However, it is also likely that there are differences in the degree of sophistication (measured by τ) across games because of an interaction between game complexity and working memory, or how well the surface game structure maps onto evolutionarily familiar games.¹²

To date, these sources of level differences have not been explored very much. Chong, Ho, and Camerer (2005) note some educational differences (Caltech students are estimated to do .5 steps of thinking more than subjects from a nearby community college) and an absence of a gender effect. Other studies have showed modest associations ($r = .3$) between strategic levels and working memory (digit span; Devetag & Warglien, 2003) and the “eyes of the mind” test of emotion detection (Georganas, Healy, & Weber, 2010).

Many papers have reported some degree of cross-game type stability in level classification. Studies that compare a choice in one game with one different game report low stability (Burchardi & Penczynski, 2010; Georganas et al., 2010). However, as is well known in personality psychology and psychometrics, intrapersonal reliability typically increases with the number of items used to construct a scale. Other studies using more game choices to classify report much higher correlations (comparable to Big 5 personality measures) (Bhui & Camerer, research in progress).

As one illustration of potential type stability, Figure 18.5 below shows estimated types for individuals

using the first 11 games in a 22-game series (x-axis) and types for the same individuals using the last 11 games. The correlation is quite high ($r = .61$). There is also a slight upward drift across the games (the average level is higher in the last 11 games compared to the first), consistent with a transfer or practice effect, even though there is no feedback during the 22 games (see also Weber, 2003).

Field Data

Since controlled experimentation came late to economics (c. 1960) compared to psychology, there is a long-standing skepticism about whether theories that work in simple lab settings generalize to naturally occurring economic activity. Five studies have applied CH or level-k modeling to auctions (Gillen, 2009), strategic thinking in managerial choices (Goldfarb & Yang, 2009; Goldfarb & Xiao, in-press), and box office reaction when movies are not shown to critics before release (Brown, Camerer, & Lovallo, in press).

One study is described here as an example (Östling et al., 2011). In 2007 the Swedish Lottery created a game in which people pay 1 euro to enter a lottery. Each paying entrant chooses an integer 1–99,999. The *lowest unique positive integer* (hence, the acronym LUPI) wins a large prize.

The symmetric equilibrium is a probabilistic profile of how often different numbers are chosen (a “mixed” equilibrium). The lowest numbers are always chosen more often (e.g., 1 is chosen most often); the rate of decline in the frequency of choice is accelerating up

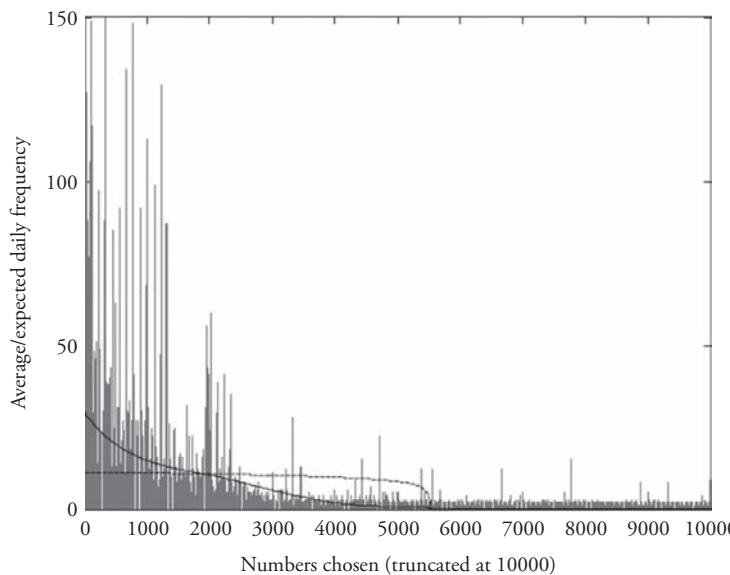


Fig. 18.6 Numbers chosen in week 1 of Swedish LUPI lottery ($N = \text{approximately } 350,000$). Dotted line indicates mixed Nash equilibrium. Solid line indicate stochastic cognitive hierarchy (CH) model with two free parameters. Best-fitting average steps of thinking is $\tau = 1.80$ and $\lambda = .0043$ (logit response).

to a sharp inflection point (number 5,513); and the rate of decline slows down after 5,513.

Figure 18.6 shows the data from only the lowest 10% of the number range, from 1–10,000 (higher number choices are rare, as the theory predicts). The predicted Nash equilibrium is shown by a dotted line—a flat “shelf” of choice probability from 1 to 5,513, then a sharp drop. A fitted version of the CH model is indicated by the solid line. CH can explain the large frequency of low number choices (below 1,500), since these correspond to low levels of strategic thinking (i.e., people don’t realize everyone else is choosing low numbers, too). Since level-0 types randomize, their behavior produces too many high numbers (above 5,000). Since the lowest and highest numbers are chosen too often according to CH, compared to the equilibrium mixture, CH also

implies a gap between predicted and actual choices in the range 2,500–5,000. This basic pattern was replicated in a lab experiment with a similar structure. While there are clear deviations from Nash equilibrium, consistent with evidence of limited strategic thinking, in our view the Nash theory prediction is not bad considering that it uses no free parameters and comes from an equation which is elegant in structure but difficult to derive and solve.

The LUPI game was played in Sweden for 49 days in a row, and results were broadcast on a nightly TV show. Analysis indicates an imitate-the-winner fictive learning process, since choices on one day move in the direction of 600-number range around the previous day’s winner. The result of this imitation is that every statistical feature of the numbers chosen moves toward the equilibrium across the 7 weeks.

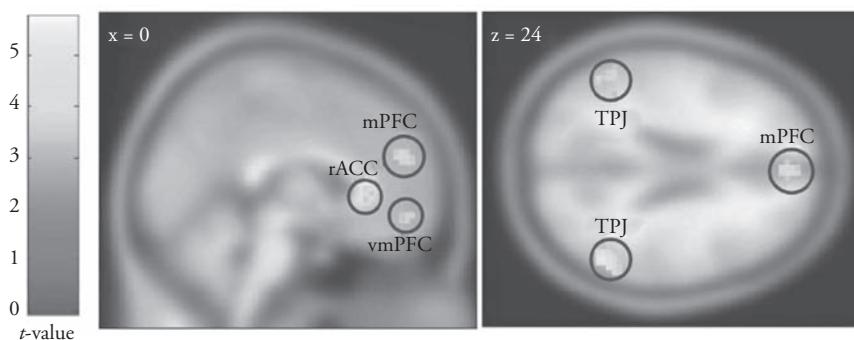


Fig. 18.7 Brain regions more active in level-2 reasoners compared to level-1 reasoners (classified by choices), differentially in playing human compared to computer opponents. mPFC, medial prefrontal cortex; rACC, rostral anterior cingulate cortex; TPJ, temporoparietal junction; vmPFC, ventromedial prefrontal cortex. (From Coricelli & Nagel, 2009, Fig. S2a.) See color figure.

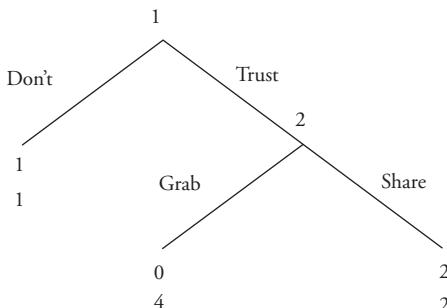


Fig. 18.8 A simple trust game (Dufwenberg & Gneezy, 2000).

For example, in the last week the average number is 2,484, within 4% of the predicted value of 2,595.

Psychological Games

In many strategic interactions, our own beliefs or beliefs of other people seem to influence how we value consequences. For example, surprising a person with a wonderful gift that is perfect for her is more fun for everyone than if the person had asked for it. Some of that pleasure comes from the surprise itself.

This type of pattern can be modeled as a “psychological game” (Geanakoplos, Pearce, & Stacchetti, 1989; Battigalli & Dufwenberg, 2009). PGs are an extension of standard games in which the utility evaluations of outcomes can depend on beliefs about what was thought to be likely to happen (as well as typical material consequences). This approach requires thinking and reasoning since the belief is derived from analysis of the other person’s motives. Together these papers provide tools for incorporating motivations such as intentions, social norms, and emotions into game-theoretic models.

Emotions are an important belief-dependent motivation. Anxiety, disappointment, elation, frustration, guilt, joy, regret, and shame, among other emotions, can all be conceived of as belief-dependent incentives or motivations and incorporated into models of behavior using tools from psychological game theory.

One example is guilt: Baumeister, Stillwell, and Heatherton (1994) write: “If people feel guilty for hurting their partners...and for failing to live up to their expectations, they will alter their behavior (to avoid guilt).” Battigalli and Dufwenberg (2007) operationalize the notion that people will anticipate and avoid guilt in their model of guilt aversion. In their model, players derive positive utility from both material payoffs and negative utility from

guilt. Players feel guilty if their behavior disappoints a coplayer relative to his expectations.¹³

Consider Figure 18.8, which illustrates a simple trust game from Dufwenberg (2002). Player 1 may choose either “Trust” or “Don’t.” In the first case player 1 gets the move, while after a choice of “Don’t” the game ends and each player gets payoff 1. If player 2 gets the move, she chooses between “Grab” and “Share.” The payoffs to Grab are 0 for player 1 and 4 for player 2.

The subgame perfect equilibrium of this game for selfish players is for Player 2 to choose Grab if she gets the move since it results in a higher payoff for her than choosing Share. Player 1 anticipates this behavior and chooses Don’t to avoid receiving 0. Both players receive a payoff of 1, which is inefficient.

Now suppose that Player 2 is guilt averse. Then her utility depends not only on her material payoff but also on how much she “lets down” player 1 relative to his expectations. Let p be the probability that player 1 assigns to “Share.” Let p' represent Player 2’s (point) belief regarding p , and suppose that 2’s payoff from Grab is then $4 - \theta p$, where theta represents player 2’s sensitivity to guilt. If player 1 chooses Trust it must be that p is greater than $\frac{1}{2}$; otherwise player 1 would choose Don’t. Then if $\theta \geq 2$, player 2 will choose Share to avoid the guilt from letting down Player 1. Knowing this, player 1 will choose Trust. In this outcome both players receive 2 (instead of 1 in the selfish subgame perfect equilibrium), illustrating how guilt aversion can foster trust and cooperation where selfish behavior leads to inefficiency.

A number of experiments have studied guilt aversion in the context of trust games, including Dufwenberg and Gneezy (2000), Charness and Dufwenberg (2006, 2011), and Reuben et al. (2009). All of these papers find evidence that a desire to avoid guilt motivates players to behave unselfishly by reciprocating trust (for a contrary opinion see Ellingsen et al., 2010). Recent fMRI evidence (Chang et al., 2011) suggests that avoiding guilt in trust games is associated with increased activity in the anterior insula.

Psychological game theory also may be employed to model other social emotions such as shame (Tadelis, 2011) or anger (Smith, 2009) or to import existing models of emotions such as disappointment, elation, regret, and rejoicing (Bell, 1982, 1985; Loomes & Sugden, 1982, 1986) into games.¹⁴ Battigalli and Dufwenberg (2009) provide some examples of these applications. These models are just a glimpse of the potential applications of

psychological game theory to the interaction of emotion and cognition in social interactions.

Another important application of psychological game theory is sociological concerns, such as reciprocity (which may be driven by emotions). In an important work, Rabin (1993) models reciprocity via functions that capture a player's "kindness" to his coplayer and the other player's kindness to him. These kindness functions depend on the players' beliefs regarding each other's actions and their beliefs about each other's beliefs. Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006) extend Rabin's model to sequential games.

Psychological game theory provides a useful toolkit for incorporating psychological, social, and cultural factors into formal models of decision making and social interactions. Many applications remain to be discovered and tested via experiment.

Conclusions

Compared to its impact on other disciplines, game theory has had less impact in cognitive psychology so far. This is likely because many of the analytical concepts used to derive predictions about human behavior do not seem to correspond closely to cognitive mechanisms. Some game theorists have also complained about this unrealism. Eric Van Damme (1999) wrote:

Without having a broad set of facts on which to theorize, there is a certain danger of spending too much time on models that are mathematically elegant, yet have little connection to actual behavior. At present our empirical knowledge is inadequate and it is an interesting question why game theorists have not turned more frequently to psychologists for information about the learning and information processes used by humans (p. 204).

But recently an approach called behavioral game theory has been developed that uses psychological ideas to explain both choices in many different games, and associated cognitive and biological (Bhatt & Camerer, 2011; Camerer, 2003).

This chapter discussed two elements of behavioral game theory that might be of most interest to cognitive psychologists: the cognitive hierarchy approach; and psychological games in which outcome values can depend on beliefs, often accompanied by emotions (e.g., a low bargaining offer could create anger if you expected more, or joy if you expected less).

The cognitive hierarchy approach assumes that some players choose rapidly and heuristically

("level 0") and higher-level players correctly anticipate what lower-level players do. The theory has been used to explain behavior in lab games which is both far from and close to equilibrium in different games, is supported by evidence from visual eye tracking and Mouselab, is evident in "theory of mind" circuitry during fMRI, and also can explain some patterns in field data (such as the Swedish LUPI lottery).

Research on psychological games is less well developed empirically but has much promise for understanding phenomena like "social image," norm enforcement, how emotions are created by surprises, and the relationship between emotion, cognition, and strategic behavior.

Future Directions

There are a lot of open research questions in which combining cognitive science and game theory would be useful. Here are a few:

1. Can the distribution of level types be derived endogenously from more basic principles of cognitive difficulty and perceived benefit, or perhaps from evolutionary constraint on working memory and theory of mind (e.g., Stahl, 1993)?
2. CH models have the potential to describe differences in skill or experience. Skill arises in everyday discussions about even simple games like rock, paper, and scissors; in games with private information such as poker; and games that tax working memory such as chess. Are skill differences general or domain-specific? Can skill be taught? How does skill development change cognition and neural activity?
3. The computational approach to strategic thinking in behavioral game theory could be useful for understanding the symptoms, etiology, and treatment of some psychiatric disorders. Disorders could be conceptualized as failures to correctly anticipate what other people do and feel in social interactions or to make good choices given sensible beliefs. For example, in repeated trust games King-Casas et al. (2008) found that patients with borderline personality disorder did not have typical activity in insula cortex in response to being mistrusted and earned less money because of the inability to maintain steady reciprocal trust behaviorally. Chiu et al. (2008) found that autism patients had less activity in a region of anterior cingulate that typically encodes signals of valuation during one's own strategic choices (compared to choices of others).

4. A small emerging approach in the study of literature focuses on the number of mental states that readers can track and their effect (e.g., Zunshine, 2006). One theory is that three mental states are a socially important reasonable number (e.g., love triangles) and are therefore narratively engaging. Work on cognitive schemas, social categorization, computational linguistics, and game theory could therefore be of interest in the study of literature.
5. Formal models connecting emotions with beliefs, actions, and payoffs can illuminate the relationships between affective states and behavior. The utility function approach to modeling emotions makes clear that emotions influence behavior only when the hedonic benefits of emotional behavior outweigh the costs. This approach, which considers even emotion-driven behavior as the outcome of an optimization problem (perhaps sculpted by human evolution rather than conscious cost-benefit, of course), promises to open up new avenues of research studying the relationship between emotion and strategic choices. These theories could also help organize interesting findings about neural correlates and biological bases of social emotions (e.g., Crockett, Clark, Tabibnia, Lieberman, & Robbins, 2008; Izuma, Saito, & Sadato, 2010; Takahashi, Kato, Matsuura, Mobbs, Suhara, & Okubo, 2009).

Notes

1. Another important component of behavioral game theory is learning from repeated play (perhaps using reinforcement rules as well as model-based “fictive learning” (Camerer & Ho, 1999). Learning models are widely studied but lie beyond the scope of this chapter (see e.g., Fudenberg & Levine, 1998; Camerer, 2003, chapter 6).

2. Other mechanisms that could produce equilibration include learning from observation, introspection, calculation (such as firms hiring consultants to advise on how to bid on auctions), imitation of attention-getting or successful strategies or people, or a process of preplay talking about future choices. The learning literature is well developed (e.g., Camerer, 2003, chapter 6), but the study of imitation and preplay talking could certainly use more collaboration between game theorists and psychologists.

3. Common knowledge requires, for two players, that A knows that B knows that A knows . . . ad infinitum.

4. A more general view is that level 0’s choose intuitively or “heuristically” (perhaps based on visually salient strategies or payoffs, or “lucky numbers”), but that topic has not been explored very much.

5. Restricting communication is not meant to be realistic and certainly is not. Instead, communication is restricted because choosing what to say is itself a “strategy” choice which complicates analysis of the game—it opens a Pandora’s box of possible

effects that lie outside the scope of standard game theory. However, game theorists are well aware of the possible powerful effects of communication and have begun to study it in simple ways. In her thesis Nagel (1995) reports some subject debriefing that is illustrative of CH thinking, and Sbriglia (2008) reports some protocols, too. Burchardi and Penczynski (2010) also used chat messaging and team choice to study communication and report evidence largely consistent with CH reasoning.

6. This analysis assumes $\tau = 1.5$, but the general point holds more widely.

7. Note that this is a close relative of a “threshold public goods” game. In that game, a public good is created, which benefits everyone, if T people contribute, but if even one person does not, the public good is not produced. In that case, everyone would like to be in the N-T group of people who benefit without paying.

8. $p(X)$ is a “common prior.” An example is a game of cards, in which everyone knows the contents of the card deck, do not know what face-down cards other players are holding but also know that the other players do know their own face-down cards.

9. There are several subtle variants. In the original Mouselab, boxes open and close automatically when the mouse enters and exits. Costa-Gomes, Crawford and Broseta (2001) wanted more deliberate attention, so they created a version in which a click is required to open a box. Brocas et al. (2009) created a version that requires the mouse button to be held down to view box contents (if the button press is halted the information disappears).

10. The rostral ACC, labeled rACC, is more active in level-1 than in level-2 players in the human-computer contrast.

11. DLPFC is also involved in cognitive regulation of emotions (e.g., Ochsner et al., 2009).

12. What we have in mind here is similar to Cheng and Holyoak (1985) and Fiddick, Cosmides, and Tooby’s (2000) arguments about the difference between abstract logic performance and contextualized performance. For example, games that resemble hiding food and guarding hidden locations might map roughly onto something like poker, whereas a lot of games constructed for challenge and entertainment, such as chess, do not have clear counterparts in ancestral adaptive environments.

13. Other models of belief-dependent utility can be placed in the general framework of Battigalli and Dufwenberg (2009). For example, Caplin and Leahy (2004) model doctor–patient interactions where uncertainty may cause patient anxiety. The doctor is concerned about the patient’s well-being and must decide whether to provide (potentially) anxiety-causing diagnostic information. Bernheim (1994) proposes a model of conformity where players care about the beliefs their coplayers have regarding their preferences. The model can produce fads and adherence to social norms. Related work by Benabou and Tirole (2006) models players who are altruistic and also care about other’s inferences about how altruistic they are. Gill and Stone (2010) model players who care about what they feel they deserve in two-player tournaments. The players’ perceived entitlements depend upon their own effort level and the efforts of others.

14. A (single person) decision problem involving any of these emotions may be modeled as a psychological game with one player and moves by nature.

Acknowledgments

This research was supported by The Betty and Gordon Moore Foundation and by National Science Foundation grant NSF-SES 0850840.

References

- Akerlof, G. A. (1970). The market for "lemons": quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84, 488–500.
- Amadio, D. M., & Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, 7, 268–277.
- Aumann, R., & Brandenburger, A. (1995). Epistemic conditions for Nash equilibrium. *Econometrica*, 63, 1161–1180.
- Baumeister, R., Stillwell, A., & Heatherton, T. (1994). Guilt: an interpersonal approach. *Psychological Bulletin*, 115, 243–267.
- Battigalli, P., & Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, 97, 170–176.
- Battigalli, P., & Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory*, 144, 1–35.
- Bazerman, M. H., & Samuelson, W. F. (1983). I won the auction but don't want the prize. *Journal of Conflict Resolution*, 27, 618–634.
- Benabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96, 1652–1678.
- Bell, D. E. (1982). Regret in decision making under uncertainty. *Operations Research*, 30, 961–981.
- Bell, D. E. (1985). Disappointment in decision making under uncertainty. *Operations Research*, 33, 1–27.
- Bernheim, B. D. (1994). A theory of conformity. *The Journal of Political Economy*, 102, 841–877.
- Bhatt, M., & Camerer, C. F. (2005). Self-referential thinking and equilibrium as states of mind in games: fMRI evidence. *Games and Economic Behavior*, 52, 424.
- Bhatt, M., & Camerer, C. F. (2011). The cognitive neuroscience of strategic thinking. In J. Cacioppo & J. Decety (Eds.), *Handbook of social neuroscience*. Oxford, England: Oxford University Press.
- Bhatt, M. A., Lohrenz, T., Camerer, C. F., & Montague, P. R. (2010). Neural signatures of strategic types in a two-person bargaining game. *Proceedings of the National Academy of Sciences*, 107(46), 19720–19725.
- Binmore, K., Shaked, A., & Sutton, J. (1985). Testing noncooperative bargaining theory: a preliminary study. *The American Economic Review*, 75, 1178–1180.
- Bosch-Domenech, A., Garcia-Montalvo, J., Nagel, R., & Satorra, A. (2002). One, two, (three), infinity...: Newspaper and lab beauty-contest experiments. *American Economic Review*, 92, 1687–1701.
- Burchardi, K. B., & Penczynski, S. P. (2010). Out of your mind: eliciting individual reasoning in one shot games. LSE working paper, April.
- Brandenburger, A. (1992). Knowledge and equilibrium in games. *Journal of Economic Perspectives*, 6(4), 83–101.
- Brocas, I., Carrillo, J. D., Wang, S., & Camerer, C. F. (2009). *Measuring attention and strategic behavior in games with private information*, Mimeo edn, Pasadena.
- Brown, A. L., Camerer, C. F., & Lovallo, D. (in press). To review or not review? Limited strategic thinking at the movie box office. *American Economic Journal: Microeconomics*.
- Camerer, C. (1990). Behavioral game theory. In R. Hogarth (Ed.), *Insights in decision making: a tribute to Hillel J. Einhorn* (pp. 311–336). Chicago: University of Chicago Press.
- Camerer, C. F. (2003). *Behavioral game theory*, Princeton: Princeton University Press.
- Camerer, C., & Ho, T. H. (1999). Experience-weighted attraction learning in normal form games. *Econometrica*, 67, 827–874.
- Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2002). Sophisticated experience-weighted attraction learning and strategic teaching in repeated games. *Journal of Economic Theory*, 104, 137–188.
- Camerer, C. F., Ho, T. H., & Chong, J. K. (2004). A cognitive hierarchy model of games, *Quarterly Journal of Economics*, 119, 861–898.
- Camerer, C. F., Johnson, E., Rymon, T., & Sen, S. (1993). Cognition and framing in sequential bargaining for gains and losses. In K. G. Binmore, A. P. Kirman, & P. Tani (Eds.), *Frontiers of game theory* (pp. 27–47). Cambridge: MIT Press.
- Capen, E. C., Clapp, R. V., & Campbell, W. M. (1971). Competitive bidding in high-risk situations. *Journal of Petroleum Technology*, 23, 641–653.
- Caplin, A., & Leahy, J. (2004). The supply of information by a concerned expert. *The Economic Journal*, 114, 487–505.
- Cavanna, A. E., & Trimble, M. R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain: A journal of neurology*, 129(3), 564–583.
- Chang, L., Smith, A., Dufwenberg, M., & Sanfey, A. (2011). Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron*, 70(3), 560–572.
- Charness, G., & Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74, 1579–1601.
- Charness, G., & Dufwenberg, M. (2011). Participation. *The American Economic Review*, 101(4), 1211–1237.
- Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, 17, 391–416.
- Chiu, P. H., Kayali, M. A., Kishida, K. T., Tomlin, D., Klinger, L. G., Klinger, M. R., & Montague, P. R. (2008). Self responses along cingulate cortex reveal quantitative neural phenotype for high-functioning autism. *Neuron*, 57, 463–473.
- Chong, J. K., Camerer, C. F., & Ho, T.-H. (2005). Cognitive hierarchy: a limited thinking theory in games. *Experimental Business Research – Volume III: Marketing, Accounting and Cognitive Perspectives* (pp. 203–228). Boston, MA: Springer.
- Chong, J., Camerer, C. F., & Ho, T. H. (2006). A learning-based model of repeated games with incomplete information. *Games and Economic Behavior*, 55, 340–371.
- Coricelli, G., & Nagel, R. (2009). Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *PNAS*, 106, 9163–9168.
- Costa-Gomes, M., Crawford, V. P., & Broseta, B. (2001). Cognition and behavior in normal-form games: An experimental study. *Econometrica*, 69, 1193–1235.
- Costa-Gomes, M. A., & Crawford, V. P. (2006). Cognition and behavior in two-person guessing games: an experimental study. *American Economic Review*, 96, 1737–1768.
- Crawford, V. P., Costa-Gomes, M. A., & Iribarri, N. (2010). Strategic thinking. Working paper.
- Crockett, M. J., Clark, L., Tabibnia, G., Lieberman, M. D., & Robbins, T. W. (2008). Serotonin modulates behavioral reactions to unfairness. *Science*, 320, 1739.
- Deverag, G., & Warglien, M. (2003). Games and phone numbers: Do short-term memory bounds affect strategic behavior? *Journal of Economic Psychology*, 24, 189–202.
- Dufwenberg, M. (2002). Marital investments, time consistency and emotions. *Journal of Economic Behavior & Organization*, 48(1), 57–69.

- Duffy, J., & Hopkins, E. (2005). Learning, information, and sorting in market entry games: theory and evidence. *Games and Economic Behavior*, 51, 31–62.
- Dufwenberg, M., & Gneezy, U. (2000). Measuring beliefs in an experimental lost wallet game. *Games and Economic Behavior*, 30, 163–182.
- Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, 47, 268–298.
- Duffy, J., & Hopkins, E. (2005). Learning, information, and sorting in market entry games: theory and evidence. *Games and Economic Behavior*, 51, 31–62.
- Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science*, 302, 290–292.
- Ellingsen, T., Johannesson, M., Tjotta, S., & GauteTorsvik, G. (2010). Testing guilt aversion. *Games and Economic Behavior*, 68, 95–107.
- Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54(2), 293–315.
- Fiddick, L., Cosmides, L., & Tooby, J. (2000). No interpretation without representation: the role of domain-specific representations and inferences in the Wason selection task. *Cognition*, 77, 1–79.
- Fudenberg, D., & Levine, D. (1998). *Theory of learning in games*. Cambridge, MA: MIT Press.
- Gallagher, H. L., Jack, A. I., Poepstorf, A., & Frith, C. D. (2002). Imaging the Intentional Stance in a Competitive Game. *NeuroImage*, 16, 814–821.
- Geanakoplos, D., Pearce, D., & Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, 1, 60–79.
- Gill, D., & Stone, R. (2010). Fairness and desert in tournaments. *Games and Economic Behavior*, 69, 346–364.
- Gillen, B. (2009). Identification and estimation of Level-k auctions. Working paper.
- Glimcher, P. W., Camerer, C. F., Fehr, E., & Poldrack, R. (Eds.) (2008). *Neuroeconomics: decision making and the brain*. London: Academic.
- Georganas, S., Healy, P. J., & Weber, R. (2010). On the persistence of strategic sophistication. Working paper.
- Goldfarb, A., & Xiao, M. (in-press). Who thinks about the competition? Managerial ability and strategic entry in US local telephone markets. *American Economic Review*.
- Goldfarb, A., & Yang, B. (2009). Are all managers created equal? *Journal of Marketing Research*, 46, 612–622.
- Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences*, 105(18), 6741–6746.
- Harsanyi, J. C. (1967). Games with incomplete information played by 'Bayesian' players, I–III. Part I. The basic model. *Management Science*, 14, 159–182.
- Hedden, T., & Zhang, J. (2002). What do you think I think you think? Theory of mind and strategic reasoning in matrix games. *Cognition*, 85, 1–36.
- Hsu, M., Anen, C., & Quartz, S. R. (2008). The right and the good: distributive justice and neural encoding of equity and efficiency. *Science*, 320, 1092–1095.
- Izuma, K., Saito, D. N., & Sadato, N. (2010). Processing of the incentive for social approval in the ventral striatum during charitable donation. *Journal of Cognitive Neuroscience*, 22, 621–631.
- Johnson, E. J., Camerer, C., Sen, S., & Rymon, T. (2002). Detecting failures of backward induction: monitoring information search in sequential bargaining. *Journal of Economic Theory*, 104, 16–47.
- Kahneman, D. (1988). Experimental economics: a psychological perspective. In R. Tietz, W. Albers, & R. Selten (Eds.), *Bounded rational behavior in experimental games and markets* (pp. 11–18). New York: Springer-Verlag.
- Keynes, J. M. (1936). *The general theory of employment, interest, and money*. London: Macmillan.
- King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *Science*, 321, 806–810.
- Kross, E., Berman, M. G., Mischel, W., Smith, E. E., & Wager, T. D. (2011). Social rejection shares somatosensory representations with physical pain. *Proceedings of the National Academy of Sciences*, 108(15), 6270–6275.
- Kuo, W. J., Sjostrom, T., Chen, Y. P., Wang, Y. H., & Huang, C. Y. (2009). Intuition and deliberation: two systems for strategizing in the brain. *Science*, 324, 519–522.
- Loomes, G., & Sugden, R. (1982). Regret theory: an alternative theory of rational choice under uncertainty. *The Economic Journal*, 92, 805–824.
- Loomes, G., & Sugden, R. (1986). Disappointment and dynamic consistency in choice under uncertainty. *The Review of Economic Studies*, 53, 271–282.
- Luce, R. D. (1959). *Individual choice behavior*. Oxford, England: John Wiley.
- McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences of the United States of America*, 98, 11832–11835.
- Nagel, R. (1995). Unraveling in guessing games: an experimental study. *The American Economic Review*, 85, 1313–1326.
- Nash, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences of the United States of America*, 36, 48–49.
- Neelin, J., Sonnenschein, H., & Spiegel, M. (1988). A further test of noncooperative bargaining theory. *American Economic Review*, 78, 824–836.
- Ochsner, K., Hughes, B., Robertson, E., Gabrieli, J., Cooper, J., & Gabrieli, J. (2009). Neural systems supporting the control of affective and cognitive conflicts. *Journal of Cognitive Neuroscience*, 21, 1841–1854.
- Östling, R., Wang, J. T.-y., Chou, E., & Camerer, C. F. (2011). Testing game theory in the field: Swedish LUPI lottery games. *American Economic Journal: Microeconomics*, 3(3), 1–33.
- Preuschoff, K., Quartz, S. R., & Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. *Journal of Neuroscience*, 28, 2745–2752.
- Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, 83, 1281–1302.
- Reuben E., Sapienza P., & Zingales L. (2009). Is mistrust self-fulfilling? *Economics Letters*, 104, 89–91.
- Rotemberg, J. J. (2008). Minimally acceptable altruism and the ultimatum game. *Journal of Economic Behavior & Organization*, 66(3–4), 457–476.
- Rubinstein, A. (1982). Perfect equilibrium in a bargaining model. *Econometrica*, 50(1), 97–110.
- Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, 300, 1755–1758.

- Sbriglia, P. (2008). Revealing the depth of reasoning in p-beauty contest games. *Experimental Economics*, 11, 107–121.
- Smith, A. (2009). Belief-dependent anger in games. Working paper.
- Stahl, D. O. (1993). Evolution of smart n players. *Games and Economic Behavior*, 5, 604–617.
- Tadelis, S. (2011). The power of shame and the rationality of trust. Working paper.
- Takahashi, H., Kato, M., Matsuura, M., Mobbs, D., Suhara, T., & Okubo, Y. (2009). When your gain is my pain and your pain is my gain: neural correlates of envy and schadenfreude. *Science*, 323, 937–939.
- Van Damme, E. (1999). Game theory: the next stage. In L. A. Gérard-Varet, A. P. Kirman, & M. Ruggiero (Eds.), *Economics beyond the millennium* (pp. 184–214). Oxford, UK: Oxford University Press.
- Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.
- Weber, R. A. (2003). ‘Learning’ with no feedback in a competitive guessing game. *Games and Economic Behavior*, 44, 134–144.
- Yoshida, W., Seymour, B., Friston, K. J., & Dolan, R. (2009). Neural mechanism of belief inference during cooperative games. *Journal of Neuroscience*, 30, 10744–10751.
- Zunshine, L. (2006). *Why we read fiction: theory of mind and the novel*. Columbus: The Ohio State University Press.

Michael R. Waldmann, Jonas Nagel, and Alex Wiegmann

Abstract

The past decade has seen a renewed interest in moral psychology. A unique feature of the present endeavor is its unprecedented interdisciplinarity. For the first time, cognitive, social, and developmental psychologists, neuroscientists, experimental philosophers, evolutionary biologists, and anthropologists collaborate to study the same or overlapping phenomena. This review focuses on moral judgments and is written from the perspective of cognitive psychologists interested in theories of the cognitive and affective processes underlying judgments in moral domains. The review will first present and discuss a variety of different theoretical and empirical approaches, including both behavioral and neuroscientific studies. We will then show how these theories can be applied to a selected number of specific research topics that have attracted particular interest in recent years, including the distinction between moral and conventional rules, moral dilemmas, the role of intention, and sacred/protected values. One overarching question we will address throughout the chapter is whether moral cognitions are distinct and special, or whether they can be subsumed under more domain-general mechanisms.

Key Words: moral psychology, moral judgment, norms, moral domains, intention, folk psychology, emotion, reasoning, cross-cultural psychology, neuroscience, trolley problem, convention, protected and sacred values, heuristics and biases, dual-process theory, moral grammar, side-effect effect

Introduction

The past decade has seen a renewed interest in moral psychology. Empirical research on morality is not new, of course. There has been a long tradition in different fields, such as social and developmental psychology. Nevertheless, a unique feature of the present endeavor is its unprecedented interdisciplinarity. For the first time, cognitive, social, and developmental psychologists, neuroscientists, experimental philosophers, evolutionary biologists, and anthropologists collaborate to study the same or overlapping phenomena.

In this review, we will focus on research trying to elucidate the cognitive and affective foundations of *moral judgment*. As a first approximation

one can say that moral judgments refer to the rightness or wrongness of specific acts or policies. A central question that will be repeatedly brought up in this review is whether we need a separate field of moral psychology to study this specific class of judgments. Do moral judgments possess characteristics that make them qualitatively distinct from other judgments? This question can be divided into two subquestions: (1) Are there moral rules people universally invoke when making moral judgments, and (2) Are moral cognitions a natural kind with specialized cognitive machinery, or are moral judgments just a special case of judgments in general? We will try to answer these two questions by reviewing recent studies investigating whether moral rules

are universal and whether there is evidence for an innate module devoted to moral cognitions.

The answer to the first question seems to be no. Cross-cultural research has made it clear that although a variety of moral rules, such as “do no harm,” are strongly endorsed in Western cultures, at least by liberals, they are not universally endorsed (see Rai & Fiske, 2011). For example, whereas in some societies hitting and fighting is impermissible, in other cultures certain forms of violence are praised. Some cultures allow violence only to outgroup members; others also encourage violence within their ingroup and find it acceptable that children, women, or animals are harmed in some circumstances. Other moral domains, such as concerns about sexuality, fairness, health, or food, are also highly variable (see Prinz, 2007; Rai & Fiske, 2011; Sripada & Stich, 2006). Thus, the contents of moral rules vary widely across cultures.

One possibility to address the second question is to look at evidence showing that moral cognitions are innate. The evidence about lack of universality already indicates that it is unlikely that specific moral rules (e.g., “do no harm”) are innate. Although some researchers argue that some moral rules may be components of a universal moral grammar (Hauser, 2006; Mikhail, 2011), the evidence for this claim is weak (see Prinz, 2007; also see section on “Moral Grammar Theory”). Some researchers have therefore proposed that moral cognitions are nothing special: Moral reasoning is just domain-general reasoning with moral contents. Bucciarelli, Khemlani, and Johnson-Laird (2008) have, for example, claimed that moral reasoning can be modeled as deontic reasoning with the contents of the rules determined by cultural norms. Similarly, Gigerenzer (2010) claims that there is no special class of moral heuristics. Instead the same domain-general social heuristics guide moral and nonmoral behavior (e.g., “If there is a default, do nothing about it”). Although there can be no doubt that domain-general processes influence moral reasoning (see section on “Domain-General Cognitive Theories”), it also seems implausible to fully reduce moral rules to deontic rules. Deontic rules do not differentiate between moral rules and mere, arbitrary conventions, which seem psychologically distinct (see section on “The Moral/Conventional Distinction”).

Currently there is a debate about whether the evidence favors the theory that we are innately disposed to acquire *moral rules*, which share a core content that may be variably instantiated in different

cultures (Hauser, 2006; Joyce, 2006), or whether we are simply disposed to acquire *norms* whatever their content may be. Sripada and Stich (2006) differentiate norms from both moral rules and mere conventions. Like moral rules, norms typically transcend mere conventions and are considered independent of what an authority says. People believe that the norms they follow should be honored as ends, not as means to achieve a goal. Moreover, norm violations often lead to punitive emotions, such as anger or guilt. However, people who endorse norms do not necessarily claim universality for them, a feature typically associated with moral rules (see section on “The Moral/Conventional Distinction”). Sripada and Stich (2006) have suggested a domain-general norm acquisition mechanism that is capable of acquiring norms, including moral ones, the specific contents of which are culture-dependent.

This dispute is hard to settle because it depends on how moral norms are distinguished from norms in general. The review by Machery and Mallon (2010) concludes that currently the most parsimonious account is that people are universally disposed to acquire norms in general. Moral norms are then a special case; their contents are specified by the culture in which a person is born.

The question whether there is a cognitive module devoted to moral cognitions is also complicated by the fact that moral concerns may be subdivided into different domains. Extending Shweder, Much, Mahapatra, and Park’s (1997) theory, Haidt and Joseph (2007) propose five moral domains that are characterized by unique adaptive challenges, contents, triggering stimuli, virtues, and emotions. In Western cultures concerns with Harm/Care and with Fairness/Reciprocity dominate. Harm/Care concerns are triggered by suffering and distress, especially by one’s kin, and are accompanied by the emotion of compassion. The Fairness/Reciprocity domain deals with cheating, cooperation, and deception and is accompanied by the emotions anger, gratitude, and guilt. However, there are further domains. Ingroup/Loyalty norms regulate group cooperation through pride and anger, whereas Authority/Respect norms control hierarchies by recruiting the emotions respect and fear. Finally, many cultures are concerned with Purity/Sanctity, which consists of norms referring to food, health, and sexuality (thus conceiving the body as sacred), often enforced through feelings of disgust. These moral values are not only needed to understand other cultures, but there are also differences within the Western culture. For example,

conservatives are more likely than liberals to embrace all these values. In contrast, Western liberals mainly emphasize Harm and Fairness/Justice-based concerns (Graham, Haidt, & Nosek, 2009; see also Wright & Baril, 2011).

Whereas Haidt and Joseph (2007) believe that these five domains correspond to adaptations that led to innately specified dispositions to acquire domain-specific moral norms, Rai and Fiske (2011) present an alternative theory that views norms as mechanisms to regulate specific types of social relations. Unity is the motive to care for and support the integrity of ingroups, Hierarchy is the motive to respect rank in social groups, Equality is the motive for balanced in-kind reciprocity, and Proportionality is the motive for rewards and punishments to be proportionate to merit and judgments to be based on a utilitarian calculus of costs and benefits (i.e., market pricing). People are simultaneously parts of several social relations so that moral norms may vary in different contexts. For example, harm may be prohibited within ingroups that rely on the implicit assumption of Equality, whereas it may be obligatory when it is proposed in the context of Hierarchy (e.g., war) or Proportionality (e.g., punishment). This theory is consistent with the assumption that some moral norms are innate, but innateness claims here are not made about the norms but rather about the universality of specific types of social relations, where moral cognition is still part of broader social-relational cognition.

Our review of research on moral judgment will start with a critical discussion of competing theories of moral judgment. We will present these theories along with selected experimental behavioral and neuroscientific studies supporting the respective theory. We will then discuss these theories in the context of a selected number of specific research topics that have attracted particular interest in recent years.

Critical Review of Global Theories

In this section we will review and discuss global theories of moral judgment. Although these theories typically are presented as general frameworks, it will turn out that the focus of the theories differs.

Kohlberg's Rationalist Theory

Kohlberg's (1981) important theory of moral development, which was inspired by Piaget's (1932) view is discussed in many current accounts as an example of a theory that views conscious moral reasoning as a central component of morality (Haidt, 2001; Hauser, 2006). Kohlberg's (1981) famous

method to study moral competencies was simple. He presented subjects (mainly children and adolescents) with dilemmas in which different moral factors conflicted. For example, in the famous *Heinz dilemma*, Heinz's dying wife can only be saved by taking a new drug that a pharmacist has developed. The production of the drug costs \$200, but the pharmacist charges \$2,000, double of what Heinz can pay. The pharmacist refuses to sell the drug cheaper so that Heinz eventually decides to break into the pharmacy and steal the drug. Kohlberg asked his subjects whether Heinz should have done this. He was primarily interested in the justifications for the answers, which he coded to reconstruct the level of moral reasoning.

Kohlberg found that children from many cultures typically move through a sequence of levels and sub-stages, although not everyone reaches the higher levels of reasoning (see also Crain, 1985). Level 1 represents *preconventional* morality. This level is characterized by an orientation toward the likely punishment or obedience toward fixed rules ("do not steal"). In Level 2, the level of *conventional morality*, typically reached in adolescence, values of family and society come into play. Here the children think that people should live up to the expectations of family and society, and be good persons. For example, Heinz's behavior might be defended as good, whereas the pharmacist might be described as greedy. Later within Level 2, subjects become more concerned with society as a whole with an emphasis on laws, respecting authorities, and performing duties to maintain social order. In Level 3, *postconventional* morality, the justifications transcend the level of norms and laws and focus on the legitimacy of the norms regulating society. In this stage violations of individual rights, such as liberty and life, may be invoked to justify behavior that breaks the law.

Kohlberg did not believe in innate factors driving moral development but rather viewed the transition between levels as driven by the opportunities afforded in everyday social interactions. Change may occur as a result of everyday role taking and perspective change fostering empathy, or it may be driven by reflections about moral situations.

DISCUSSION

Kohlberg was a rationalist. He believed that our moral judgments are driven by reasoning processes, and that progress in moral development is driven by reflections and discussions. Many current theories criticize this rationalist assumption. For

example, Haidt (2001) argues that moral intuitions are primarily based on unconscious intuitions, with justifications being post hoc rationalizations (see section on “Haidt’s Social Intuitionist Model”). Thus, it can be questioned whether the justifications Kohlberg elicited really caused the intuitions in the moral dilemmas. Other researchers acknowledge that occasionally moral intuitions may be based on reasoning processes, but they argue there are also important cases in which we do not have conscious access to the factors driving our intuitions (see Cushman, Young, & Hauser, 2006). A related critique is that Kohlberg’s focus on levels and stages underestimates the context dependency and variability of moral reasoning. Moral intuitions in different cases may be driven by different context factors so that the reduction to a general level may be an oversimplification.

Kohlberg’s theory has also been criticized as culturally biased (see also Gilligan, 1982, for the claim that Kohlberg’s higher levels are biased toward the reasoning of males). Kohlberg argues that people in all cultures go through the same levels, but there may be differences in the rates of development and the end state. For example, he found that in urban contexts people typically reach Level 2 and some lower stages of Level 3, whereas in tribal communities and small villages, Level 1 is rarely surpassed. Simpson (1974) argues that Kohlberg has developed a stage model based on the Western philosophical tradition and has then imposed it on non-Western cultures. Kohlberg’s response was that his theory is not about the specific values different cultures endorse but about general modes of reasoning, but this position has become increasingly questionable in light of the diminished role of justifications as evidence for morality in current theories.

Emotion-based Theories

Kohlberg’s (1981) theory can be traced back to rationalist philosophy. Moral reasoning is described as conscious deliberation; the sequence of moral stages seems to lead toward ethical positions that have been elaborated in Kant’s (1959) and Rawls’ (1971) philosophy. Many current theories are instead influenced by Hume’s (1960) philosophy of morality. Hume held the view that moral distinctions are not results of reasoning processes but can be derived, analogous to aesthetic judgments, from affectively laden moral sentiments: feelings of approval and disapproval felt by spectators who contemplate a character’s trait or action.

HAIDT’S SOCIAL INTUITIONIST MODEL

Inspired by Hume, Haidt (2001) defines moral judgments as “evaluations (good vs. bad) of the actions or character of a person that are made with respect to a set of virtues held to be obligatory by a culture or subculture” (p. 817). In his model an important distinction is between reasoning, a conscious activity in which conclusions are derived through several steps, and intuition, also a cognitive process that is characterized “as the sudden appearance in consciousness of a moral judgment, including an affective valence (good-bad, like-dislike), without any conscious awareness of having gone through steps of searching, weighing evidence, or inferring a conclusion” (p. 818). Thus, whereas reasoning is largely conscious, intuition is based on automatic, unconscious processes.

In his social intuitionist model, the primary link underwriting moral judgments is the link between the eliciting situations and moral intuitions. Reasoning processes may modify judgments, but in the model they are optional and start after initial intuitions have been formed. The role of reasoning is often to provide post hoc rationalizations of the already formed moral intuitions. Occasionally, private reasoning may override the initial intuitions, but this is relatively rare. Apart from these processes that occur within the individual, the model also contains links to other members of the social group. Other people may be influenced by moral judgments, or they may change their minds on the basis of discussions. Moreover, the intuitions and reasoning of others may influence the moral intuitions of the individual. Thus, individuals are embedded in social contexts and their norms.

Evidence for the existence of a direct intuitive link comes from a number of studies (see Haidt, 2001, 2007; Haidt & Kesebir, 2010). In several studies about harmless taboo violations (e.g., eating a pet dog; a consensual incestuous relation with birth control), many subjects judged that the acts were morally wrong but were unable to provide reasons for their judgments (i.e., moral dumbfounding; see Haidt & Hersh, 2001; Haidt, Koller, & Dias, 1993). Moral dumbfounding can be viewed as evidence for the unconscious elicitation of moral intuitions.

The automaticity of moral judgments may lead to misattributions of consciously accessible affects. For example, Wheatley and Haidt (2005) selected highly hypnotizable subjects who were given a post-hypnotic suggestion to feel a flash of disgust when

they read an arbitrary word. Moral vignettes were presented that did or did not contain the word. The results showed that moral judgments can be made more severe by the presence of the hypnotically triggered disgust. In a related study, Schnall, Haidt, Clore, and Jordan (2008) manipulated the context in which subjects made moral judgments about a character in a story. Subjects who scored high on a *private body consciousness* scale made harsher judgments in the presence of a bad smell (“fart spray”) than in its absence. Similarly, Eskine, Kacinik, and Prinz (2011) showed that the taste of a beverage influences moral judgments about other people.

Although our focus in the present chapter is on theories of moral reasoning, it should be noted that Haidt is a social psychologist who is mainly interested in processes that go beyond individual reasoning (see Haidt, 2007; Haidt & Kesebir, 2010). Individuals are embedded in large social contexts in which they influence others, as they are influenced by others. Thus, whereas many researchers have used the model of a lonely “intuitive scientist” to study moral reasoning, Haidt prefers the metaphor of an “intuitive politician.” The focus on larger social contexts also highlights the important role of group norms, cooperation, and methods of the society to punish defectors, which are often neglected in research centered on judgment and decision making (see also Haidt’s theory of moral domains in the “Introduction” to this chapter).

Discussion

Although Haidt’s primary interest is social and culture psychology, we will focus here on his ideas about moral judgment. His approach proved a valuable contrast to the rationalist approaches of Kohlberg (1981). Numerous findings show that moral judgments can be intuitive and automatic. However, from the viewpoint of cognitive psychology, the contrast between reasoning and intuition excludes important possibilities in the middle. In modern theories of reasoning, it is rarely assumed that the steps undertaken by the reasoner are fully accessible to consciousness (see Harman, Mason, & Sinnott-Armstrong, 2010). Mental model theory, Bayesian theories, and even mental logic theory postulate various processes that work below the threshold of conscious awareness.

A further problem is that the step between eliciting situation and intuition remains largely opaque in Haidt’s (2001) theory. Although Haidt acknowledges that intuitions are based on cognitive processes,

and Haidt and Kesebir (2010) even mention that heuristics may play an important role, there is no worked-out theory of how specific situations lead to particular moral intuitions. A cognitive-affective theory of moral intuitions needs to specify how moral scenarios are perceived and categorized, and how these initial appraisals are further processed. Of course, the information processing steps and representations leading to moral intuitions may well be unconscious or only partly conscious.

Finally, the claim that moral intuition is primary and that reasoning is secondary has led to critiques. Although it seems plausible that this relation often holds given that intuitions are based on faster processes than reasoning, there are certainly also cases in which people do not have clear initial intuitions and arrive at their judgments after careful deliberation (see Haidt & Kesebir, 2010; Paxton & Greene, 2010).

THE PLACE OF EMOTIONS

IN MORAL JUDGMENTS

In Haidt’s (2001) theory, affective evaluations play an important role in moral intuitions, but the exact role of them is left open. In fact, virtually every theory of moral reasoning acknowledges that emotions are an important part of our moral judgments. Even Kant (1959), in his rationalist philosophy of morals, claims that moral judgments are typically accompanied by moral feelings. What is debated is the exact place of emotions in moral judgments.

Different positions can be distinguished (see Hauser, 2006; Huebner, Dwyer, & Hauser, 2009; Prinz & Nichols, 2010). The Kantian approach, which postulates that deliberate conscious reasoning processes generally precede emotions, is refuted by the findings discussed in the section on “Haidt’s Social Intuitionist Model.” Numerous studies have shown that moral judgments are often immediately triggered without extensive reflections.

A second possibility, in the tradition of Hume’s ideas, views moral judgments as caused by distinct prior emotions. The problem with this approach is that it is unclear which emotions trigger distinctly moral judgments and how these emotions are caused. For example, feelings of disgust may alter moral evaluations (Schnall et al., 2008; Wheatley & Haidt, 2005), but not all feelings of disgust lead to moral judgments. Also there is no clear unambiguous relation between affects and judgments. Feelings of pity and compassion may occur when we observe immoral torture but also when we watch a lifesaving amputation of a leg. Thus, it seems more likely that

affects, such as disgust, moderate moral evaluations that are already independently triggered by signaling the degree of aversiveness or disutility.

These problems have led Nichols (2004) to his “sentimental rules” theory. Nichols argues that moral judgments are fed by two components: a normative theory and a system of emotions. The normative theory specifies the content of moral rules that are acquired in a specific sociocultural environment; the emotions alter their character, which includes being considered serious and authority independent (see section on “The Moral/Conventional Distinction”). Due to the presence of emotions, moral rules acquire their force and impact (i.e., emotion-backed rules). Moral judgments in the absence of emotions are possible but rare in healthy subjects. Such judgments would also not have the same force and strength as judgments based on emotion-backed rules. Prinz (2007) proposes a related theory but questions the possibility of separating emotions from moral judgments. In his view, moral concepts, such as “moral” or “immoral,” contain emotions as essential components (as in Hume’s account).

The present research does not allow us to empirically decide between these positions (see Huebner et al., 2009). Studies in which affect was manipulated prior to or simultaneous with the scenarios (e.g., Valdesolo & DeSteno, 2006; Wheatley & Haidt, 2005) show that affect influences judgments; they do not demonstrate that emotions are necessary for moral judgments. They might instead affect the interpretation of the scenario, the evaluation of the outcomes, or the interpretation of the test question.

The results of neuroimaging and neuropsychological studies are also ambiguous. Neuroimaging studies have shown that emotional responses are integral components of moral reasoning. For example, an increased activity in the frontal polar cortex (FPC) and medial frontal gyrus was seen in moral judgments compared to judgments of nonmoral claims (Moll et al., 2002; Moll, de Oliveira-Souza, & Eslinger, 2003; see also Greene, Sommerville, Nystrom, Darley, & Cohen, 2001). Similarly, neuropsychological studies provide strong evidence for the role of emotions in morality. For example, frontotemporal dementia patients, who suffer from the deterioration of their prefrontal and anterior temporal cortex, show blunted emotions, disregard for others, and a willingness to engage in moral transgressions (Damasio, 1994; Mendez, Anderson, & Shapira, 2005). However, the exact functional role of emotion remains unclear.

The most interesting evidence comes from a study by Koenigs et al. (2007; see also Ciaramelli, Muccioli, Lådavas, & di Pellegrino, 2007). Damage to the ventromedial prefrontal cortex (VMPC) leads to an emotional flattening and to a decreased ability to anticipate rewards and punishments (Damasio, 1994). Koenigs and colleagues showed that VMPC patients were generally able to evaluate moral dilemmas in which a victim needs to be sacrificed to save others like healthy subjects, but they differed in high-conflict dilemmas, which cause strong emotional responses in healthy subjects. Whereas these healthy subjects were primarily led by their affective responses, the VMPC patients opted for a consequentialist resolution, which simply compared numbers of victims (see also sections on “Dual-Process Theory,” “Moral Grammar Theory,” and “Moral Dilemmas”). Recently, Bartels and Pizarro (2011) extended these findings by showing that healthy participants who had higher scores on measures of Machiavellianism, psychopathy, and life meaninglessness indicated greater endorsement of utilitarian solutions. Again these studies show that in healthy subjects emotions influence judgments, but it is unclear how. Huebner and colleagues (2009) suggest that emotions in these cases may influence the interpretation of the scenario. Emotionally salient outcomes may be downplayed by the VMPC patients, and this in turn affects the moral evaluations.

Another strategy to elucidate the role of emotions in moral judgments is to take a closer look at the emotions that accompany judgments (see also Prinz & Nichols, 2010). The philosopher Williams (1985) has distinguished “thin” (e.g., good, bad) from “thick” moral concepts that are loaded with content (e.g., cruelty, courage). Similarly, in emotion research one can study thin affects (e.g., positive, negative) that may occur in the absence of awareness of the triggering conditions, or emotions can be described as thick relational concepts that have moral content and trigger specific moral behavior. This position views emotions as cognitively entrenched (Lazarus, 1991), and it is consistent with the views that moral emotions are constitutive of moral judgments (Prinz, 2007) or are strongly attached to moral rules (Nichols, 2004). Examples of moral emotions are anger and guilt. According to Prinz and Nichols (2010), anger is typically elicited by a violation of somebody’s autonomy and often motivates retaliatory acts. In contrast, guilt is elicited by the feeling of direct or indirect responsibility for somebody’s harm, especially when the harmed

person belongs to the ingroup so that there is a threat of separation or exclusion. Both emotions not only express affects but are constitutive for the expression of specific moral values.

Dual-Process Theory

All theories we have discussed so far acknowledge that both conscious reasoning and emotion-based intuitions play an important role in moral judgments. They differ, however, in what process they consider primary. A further theory, the dual-process model proposed by Greene and colleagues, claims that our brains contain multiple systems driving moral intuitions, one devoted to rational, the other to emotional processes. The system underlying rational deliberations is slow, effortful, and controlled, whereas the affective system consists of automatic, largely unconscious, intuition-based processes (see Cushman, Young, & Greene, 2010, for an overview).

Initial evidence for the dual-process theory comes from a neuroimaging study by Greene et al. (2001). This study investigated moral dilemmas, such as the trolley dilemma (Foot, 1967; Thomson, 1985), which will be more extensively discussed in the section on "Moral Dilemmas." The basic scenario in trolley dilemmas describes a runaway trolley threatening to kill five people on the track. Many people find it permissible to flip a switch that redirects the train onto a side track where only one person would be killed (bystander version), whereas it is generally considered impermissible to shove a person onto the tracks to stop the train (footbridge version), despite the fact that in both scenarios one person would be killed instead of five (Hauser, Cushman, Young, Jin, & Mikhail, 2007; see also Fig. 19.1). Greene and colleagues argued that the footbridge case is more *personal* and therefore triggers a negative affective response, consistent with deontological philosophy, whereas the bystander version is more *impersonal*, which therefore leads to a consequentialist weighing of lives harmed. Thus, the theory is that there are two separate systems in the brain, one triggering affect-based, and the other rational, consequentialist responses. The nature of the moral dilemma decides which system predominates. Supporting this theory, the results of Greene et al. (2001) indeed showed that brain regions associated with emotions, such as the medial prefrontal cortex, were more active with personal dilemmas, whereas brain regions associated with controlled cognitive processes such as working memory and abstract reasoning were more

active with impersonal dilemmas. Greene et al. (2001) also reported reaction time data favoring their theory, but a reanalysis by McGuire, Langdon, Coltheart, and Mackenzie (2009) shows that outliers that needed to be removed are mainly responsible for the observed pattern.

In follow-up studies, Greene and colleagues presented particularly difficult "high-conflict" dilemmas (see also "The Place of Emotions in Moral Judgments"). For example, in one example a situation is presented in which the father of a crying baby can only save his other children from enemy soldiers if he smothers his crying child to death; otherwise all children including the crying one would be killed. Greene, Nystrom, Engell, Darley, and Cohen (2004) have shown that in such cases the anterior cingulate cortex (ACC) is active, which is involved when incompatible responses are activated. Moreover, consistent with the dual-process theory, the dorsolateral prefrontal cortex was more active when what Greene et al. viewed as the "consequentialist" response (i.e., killing the baby) was given. In a related study, Greene, Morelli, Lowenberg, Nystrom, and Cohen (2008) presented subjects with such hard moral dilemmas while at the same time exposing them to a cognitively demanding secondary task. The results showed that only the "consequentialist" responses were slowed down by this procedure, which had no effect on the affect-based (according to Greene et al. "deontological") responses. Similarly Suter and Hertwig (2011) manipulated how much time they granted their subjects to render a judgment in order to constrain controlled processes. They found more "deontological" responses under strict time constraints than in the contrasting condition in which subjects had more time to respond.

Whereas these findings showed selective interference with the assumed consequentialist brain area, the studies with patients with VMPC lesions discussed in "The Place of Emotions in Moral Judgments" are evidence for interference with the emotional areas. These patients tended to give the "consequentialist" answer in high-conflict dilemmas, while the emotion-based response was blunted (Koenigs et al., 2007; Mendez et al., 2005).

DISCUSSION

Multisystem theories are attractive in many areas because they integrate a large body of different findings compared to single-system theories, which are less flexible. However, the particular version of a

dual-system theory by Greene and colleagues has drawn a number of critiques. Prinz (2008) argues that the results of Greene et al.'s (2001) study are consistent with the view that the dilemmas trigger a strong negative emotional response to harming an innocent bystander and a weaker positive emotional response to saving five potential victims. Differences in salience of the harming versus the saving options may explain the observed patterns. Moll and De Oliveira-Souza (2007) suggest that differential activations of brain areas underlie different prosocial emotions, which integrate emotional and cognitive processes rather than putting them in conflict.

A further possible critique concerns the interpretation of the contents of the two postulated brain areas. Although the exact characterization shifts across publications (see Cushman et al., 2010), the brain areas identified as underlying moral judgments were often characterized using labels describing philosophical positions, such as deontological and consequentialist (see Greene, 2008). However, it is questionable whether different brain areas embody complex contentful moral philosophies rather than more domain-general processes, such as affective reactions versus rational deliberations. Moreover, Kahane et al. (in press) have pointed out a confounding: In the experimental scenarios used by Greene and colleagues (2001, 2004) the deontological option is always also the more intuitive one. Thus, the discovered asymmetries between responses corresponding to deontological vs. consequentialist rules might in fact be due to differences in intuitiveness. Using additional scenarios in which the consequentialist response is more intuitive than the deontological one (e.g., lying in order to prevent serious harm), Kahane and colleagues demonstrated that characteristic differences in neural activation are more closely related to the intuitiveness of the response options than to their deontological versus consequentialist content.

Furthermore, the characterization of responses based on numbers of saved lives as consequentialist overstates the finding. For example, the people who judge killing the baby to be acceptable in the crying baby dilemma outlined earlier need not have applied a consequentialist moral philosophy; the same conclusion could have been reached by application of a deontological rule (Kamm, 2007) or without reference to any formal moral theory (see also Kahane & Shackel, 2008). So far, there is no evidence that a version of an intuitive consequentialist theory is coded anywhere in the brains of naïve subjects. It

would be more parsimonious to say that in different conditions different aspects of the scenarios (for example, acts versus number of victims) are highlighted (cf. Bartels, 2008). This interpretation would also have the advantage that VMPC patients and psychopaths would not have to be viewed as particularly rationalist, consequentialist reasoners (see also Bartels & Pizarro, 2011).

Finally, there are different versions of dual-process theories, which seem equally consistent with the data (see Evans, 2007; Evans, Chapter 8). One possibility is the theory endorsed by Greene and colleagues, which assumes there are two dissociable systems that operate independently, acquire different knowledge, and compete in the control of behavior. However, another possibility is that there is only a single database (e.g., trolley dilemmas with various features, such as acts and outcomes) and sequential processes operating on these representations. The initial fast processes may lead to heuristic or emotion-based judgments, whereas in some circumstances the output of the initial pass is further processed by more effortful, controlled processes (Evans, 2007; Kahneman & Frederick, 2005). This theory would also explain the findings without the need to postulate multiple brain areas embodying different moral philosophies.

Moral Grammar Theory

The theories we have discussed so far largely focus on whether moral reasoning is driven by intuitive affective processes or by conscious reasoning. None of the theories specifies a precise computational mechanism that translates situational input into moral judgments. It is the contribution of Mikhail (2007, 2011) to rise to the challenge and present a sketch of such a computational theory (see also Hauser, 2006). Like the other modern theorists, Mikhail accepts that moral judgments are typically not based on conscious deliberate reasoning. However, this does not mean that the underlying processes cannot be reconstructed as steps of computational information processing. Most cognitive theories, in both higher and lower order cognition, assume the operation of unconscious processes underlying the mental products that rise to consciousness.

Mikhail's theory of universal moral grammar is inspired by Chomsky's (1957) linguistic grammar theory. We have intuitions about the grammaticality of sentences, which can be explained as the output of the operation of a complex unconscious system of syntactic rules. Similarly, Mikhail argues that our judgments of moral permissibility may be driven by

an unconscious moral grammar that contains moral rules. The moral grammar theory holds that individuals are intuitive lawyers who possess unconscious knowledge of a rich set of legal rules along with a natural readiness to compute mental representations of human acts and omissions in legally cognizable terms. Following Chomsky, Mikhail claims that the moral grammar is innate and universal. The innateness claim is defended by a variant of the *poverty of stimulus argument*, according to which the learning input of children would not sufficiently constrain the moral rules they seamlessly acquire. A further argument is that people often have clear moral intuitions without being able to verbally explicate or justify them (Cushman et al., 2006). Empirical support for the theory mainly comes from a large Internet study in which thousands of subjects from various countries were confronted with variations of the trolley and other dilemmas (see section on “Moral Dilemmas”).

Moral grammar theory specifies a series of computational steps transforming the observed stimuli into morally relevant internal representations. Initially a set of conversion rules encodes the temporal structure of the presented stimulus (e.g., a trolley dilemma) and transforms it into a representation of the underlying causal structure. For example, in the bystander dilemma (see sections on “Dual-Process Theory” and “Moral Dilemmas”; see also Fig. 19.1), the temporally ordered events “throwing a switch,” “turning the train,” and “killing one man” are integrated into a causal chain representation. This way knowledge is acquired about morally relevant causal features, such as whether death is a side effect or a means of the proposed act or omission. Next, other conversion rules translate the causal representation into a moral representation by assigning evaluations to the effects (good vs. bad). This representation is further converted into a representation of the underlying intentional structure. In a scenario with both good and bad effects, it is by default assumed that the good outcome is the intended outcome, whereas the bad effect is a merely foreseen side effect. However, if the bad effect is a means to a good outcome, it necessarily is intended, because it constitutes the only route to the good effect. Further morally relevant information is filled in, such as whether the act is a case of intentional battery or whether the victim is harmed without having given consent. Finally, a set of deontic rules is applied to the final representation of the stimulus, yielding a judgment of obligation, permissibility or prohibition of the encoded action.

So far Mikhail (2011) has focused on two deontic rules, which are particularly relevant for trolley dilemmas. The “prohibition of battery and homicide” forbids an agent to purposely cause harm to a non-consenting victim. A second rule, the *doctrine of double effect* (DDE), can be traced back to Aquinas and to Roman Catholic theology from the 19th century. The correct interpretation of this doctrine is under dispute (see Woodward, 2001). Double effect refers to the two effects an action might have, the intended goal and a foreseen but unintended side effect. According to Mikhail’s (2009) reading of the DDE, an otherwise prohibited act, such as battery or homicide, with good and bad effects may be permissible if the prohibited act itself is not directly intended, only the good outcomes are intended, the bad ones merely foreseen, the good effects outweigh the bad one, and there are no better alternatives. This rule is consistent with the finding that people typically consider it acceptable to redirect the trolley in the bystander version but oppose the act in the footbridge version. In the bystander version the act generates a bad effect as a side effect, which is not intended but only foreseen, whereas in the footbridge version a person is directly killed as a means to a greater good. Thus, this is a case of intentional battery.

DISCUSSION

Although to date the computational theory underlying moral grammar theory is only a sketch of a model, its precision and detail vastly surpass what other moral theories currently offer. The focus on processing details may also be the reason why the scope of the model is thus far limited. It is clearly developed to account for trolley dilemma intuitions, whereas it is less clear how other moral cases will be handled.

The focus on a restricted class of harm-based dilemmas and on deontic rules that are taken from Western moral philosophy (e.g., DDE) cast doubt on the claim that the theory is universally valid as claimed. Although the Internet study has collected data in numerous cultures, we will show in the section on “Moral Dilemmas” that alternative explanations of the effects are plausible. It seems questionable that a principle, such as the DDE, is universally valid. Even an initially plausible deontic principle, such as the prohibition of intentional battery, does not seem to hold universally (see “Introduction”).

In some versions of the moral grammar theory, the analogy to Chomsky’s (1957) grammar theory is carried even further to accommodate findings of intercultural differences. In his principles and parameter

theories, Chomsky claimed that we are born with a universal innate grammar that contains fixed principles but also parameters that are set by the linguistic environment of the person. This model explains why people are able to quickly learn very different languages with differences in the syntax. Analogously, it has been argued that moral grammar may contain principles, such as the doctrine of double effect, and parameters that are set by the culture (Dwyer, 2006; Hauser, 2006; Roedder & Harman, 2010). However, no formal version of a moral grammar of this kind has been worked out yet, so this proposal remains untestable. Moreover, moral principles, such as the doctrine of double effect, combine domain-specific rules with domain-general processes (e.g., intentional and causal analyses) so that it is unclear how both types of processes are organized within an innate module (see also Cushman & Young, 2011).

There are further critiques casting doubts on the analogy between moral grammar and Chomsky's (1957) syntax theory (see also Dupoux & Jacob, 2007). First, it seems questionable to compare grammaticality judgments with permissibility judgments. Whereas with sufficient training about the meaning of the concept of syntax we know whether a sentence like "colorless green ideas sleep furiously" is grammatical, our moral intuitions, even with professional training, are extremely context sensitive and hardly ever clear cut. It seems unlikely that the factors influencing moral judgments are encapsulated in a way that warrants the modularity assumption. Moral intuitions seem to be closer to semantics and pragmatics than syntax. Moreover, the fact that people do not have conscious knowledge about moral rules does not entail innateness. There is a large literature on artificial grammar learning, for example, which similarly demonstrates judgments in the absence of valid verbal justifications (see Litman & Reber, 2005, for an overview).

These critiques do not diminish the contribution of Mikhail (2011). It is possible to work on a theory of moral rules without accepting the innateness or universality claims. In fact, the hypothesis that moral judgments are driven by moral principles, such as the doctrine of double effect, can easily be isolated from other claims (see Cushman et al., 2010) and can be tested independently (see section on "Moral Dilemmas").

Moral Heuristics

A further approach to studying moral judgments is motivated by the *heuristics and biases* paradigm,

which comes in several, often competing variants (see Sinnott-Armstrong, Young, & Cushman, 2010). A general assumption underlying research on heuristics is that people use mental shortcuts or rules of thumb that generally work well but also lead to systematic errors in specific circumstances (see Sunstein, 2005). Heuristics may operate consciously or unconsciously. We will restrict our discussion here to contentful heuristics; a discussion of the role of affect, which has also sometimes been described in terms of heuristics, will not be repeated here (see section on "The Place of Emotions in Moral Judgments").

One specific characterization of the concept of heuristics describes their use as *attribute substitution* (Kahneman & Frederick, 2005). Often target attributes T of an object X are not easily accessible so that a person instead uses an attribute H, the heuristic attribute, which is correlated with X and is more accessible. The user of the heuristic tends to believe in T when H is present. This simple model applies to many cases within the heuristics and biases program. For example, it has been shown that availability (H) is used to infer probability (T) of an event (X) (Kahneman, Slovic, & Tversky, 1982). Or in a competing theory context it has been shown that recognition (H) is often used as the basis of decisions about the size (T) of cities (X) (Gigerenzer, Todd, & the ABC Research Group, 1999).

This general framework has also been applied to moral reasoning. Baron (1994, 1998) has argued that consequentialism or utilitarianism provides normatively correct answers in questions of morality. But people who are not philosophically trained do not think along the lines of these normative theories, but rather use simple heuristics that often mislead them because they lead to overgeneralizations beyond the contexts in which they provide useful advice. Baron and his colleagues have tried to show that people are not thinking according to the normative consequentialist guidelines but instead use simple heuristics.

An example of such a heuristic is the "do no harm" heuristic, which may underlie the common intuition that it is worse for a physician to kill a patient with a deadly disease than let him die by refraining from any kind of medical intervention. Consequentialist philosophers argue that these cases should be treated equivalently (Singer, 1979). Baron and colleagues have shown in a number of well-controlled experiments that people consider harmful acts worse than harmful omissions with otherwise identical, predictable outcomes (i.e., omission bias).

For example, Spranca, Minsk, and Baron (1991) found that people find it worse when somebody who wants to harm a person offers this person a food item with an allergenic ingredient than when she passively watches the person who does not know about the ingredient taking this item himself.

A related research view has been suggested by Sunstein (2005), who subscribes to a weak non-utilitarian form of consequentialism that also might count types of acts or violations of rights as relevant consequences that need to be weighed. Otherwise the general approach is similar to Baron's. Sunstein has developed a catalog of heuristics, for example, "do no harm," "people should not engage in wrongdoing for a fee," "punish and do not reward betrayals of trust," or "do not tamper with nature" (see also Baron, 1998). This list of heuristics seems to come straight from Western deontological philosophy.

DISCUSSION

Many of the previously discussed moral theories have left the aspect of cognitive appraisal of the situation unspecified. Theories of moral heuristics represent an important step in the direction of specifying the rules that may underlie moral evaluations. However, it can be questioned whether the normative foundation of the heuristics approach holds in moral reasoning. In nonmoral tasks, such as the estimation of city sizes, the target attributes can be clearly measured and compared with the output of the heuristics. In moral domains it is far less clear what the target attributes are (see Sinnott-Armstrong et al., 2010). To evaluate a heuristic, it would be necessary to use a normative theory, and in ethics, even more than in other fields, there is no agreement about the proper normative theory. For various reasons, in psychology, consequentialism has been proposed as the yardstick for ethical judgments (Baron, 1994; Greene, 2008; Sunstein, 2005), but once we delve into the philosophical literature it becomes clear that there are various versions of consequentialist and nonconsequentialist ethics that are defensible (see Kamm, 2007; Parfit, 2011; Scanlon, 1999). For example, it is far from clear whether killing should really be viewed as equivalent to letting die (Kamm, 2007). Moreover, it can be argued that although a utilitarian cost-benefit analysis may be appropriate in small worlds with limited options, in realistic scenarios relevant information about possible outcomes, probabilities, and costs and benefits is simply not available to make such a complex strategy

reliably applicable (Bennis, Medin, & Bartels, 2010; Binmore, 2008; Gigerenzer, 2010).

Once we abandon the commitment to a specific normative ethical position, the distinction between heuristic and target attribute breaks down. A heuristic, such as "do no harm," may then be better framed as part of a deontological target theory that people happen to endorse (Sinnott-Armstrong et al., 2010). Instead of classifying simple moral rules as heuristics for a target attribute, it may be more productive to empirically study them as building blocks underlying moral judgments. It will probably turn out that one-sentence rules are too simple to explain the many subtle context effects that are known. For example, the research on the trolley dilemma shows that people do not generally invoke a "do no harm" rule; and even a more complex rule, such as the doctrine of double effect, does not provide a complete theory (see section on "Moral Dilemmas"). The research shows that far more complex intuitive theories underlie appraisal processes than the heuristics approach suggests. Moreover, complex intuitive systems of rules may only be one possible way to represent moral knowledge. Other possibilities include memory for exemplars (e.g., of moral transgressions) or prototype representations (see Harman et al., 2010; Sripada & Stich, 2006).

Domain-General Cognitive Theories

We have discussed several theoretical approaches that model the cognitive and emotional factors underlying moral judgments. However, another possible research strategy is to treat moral judgments simply as a special case of domain-general cognitive processes. For example, various researchers have shown that principles found in behavioral economics and psychological judgment and decision theory (JDM) also affect intuitions about moral scenarios (Rai & Holyoak, 2010; Reyna & Casillas, 2009). Examples include framing effects (Bartels & Medin, 2007; Kern & Chugh, 2009; Petrinovich & O'Neill, 1996; Sinnott-Armstrong, 2008), outcome bias (Gino, Shu, & Bazerman, 2010), or effects of joint versus separate evaluation (Bartels, 2008; Lombrozo, 2009; Paharia, Kassam, Greene, & Bazerman, 2009).

Another promising class of theories applicable to moral reasoning is causal model theory. Moral judgments are generally concerned with the evaluation of acts that lead to direct and indirect effects. It has been shown that the locus of intervention, the intentions causing the acts, and the causal structure leading

from acts to good and bad effects affect judgments (Cushman & Young, 2011; Sloman, Fernbach, & Ewing, 2009; Waldmann & Dieterich, 2007).

Others have pointed out the impact of attentional processes in moral judgment. As Bartels and Medin (2007; Bartels, 2008; Sachdeva & Medin, 2008) have demonstrated, moral scenarios may be evaluated very differently depending on where subjects' attention is directed (see also section on "Sacred/Protected Values"). In a related vein, Waldmann and Wiegmann (2010) argued that aspects of the causal structure of moral dilemmas may affect moral judgment by influencing people's attentional focus on alternative counterfactual contrasts (see section on "Moral Dilemmas").

The list of domain-general factors influencing moral judgment is much longer still. Abstract, high-level construal of actions leads to amplified ascriptions of both blame and praise to agents compared to more low-level, concrete representations of the same actions (Eyal, Liberman, & Trope, 2008). Metacognitive experiences like processing fluency are also used as input for moral judgment (Laham, Alter, & Goodwin, 2009; Rai & Holyoak, 2010). Approach- and avoidance-based motivational systems seem to have similar effects on judgments about moral issues as they do in other domains (Janoff-Bulman, Sheikh, & Hepp, 2009). Extensive mental simulation, induced by closed eyes during the judgment process, makes moral judgments more extreme (Caruso & Gino, 2011). And recently, more and more researchers have pointed out the importance of individual differences in moral judgment, including need for cognition (Bartels, 2008; Bartels & Pizarro, 2011), working memory capacity (Moore, Clark, & Kane, 2008), sensitivity to reward and punishment (Moore, Stevens, & Conway, 2011), and personality traits such as extraversion (Feltz & Cokely, 2009).

Specific Research Topics

After having outlined the main theoretical approaches to the study of moral judgment, along with relevant empirical evidence taken as support of them, we now turn to four selected empirical research areas that have attracted much attention in the recent years. The aim of this section is to demonstrate the complexity of explaining moral judgments in specific tasks.

The Moral/Conventional Distinction

One of the central controversies in the field of moral psychology concerns the question of whether

morality constitutes an independent domain with specific norms. Do humans reason qualitatively differently about moral rules as opposed to mere social conventions? This question has motivated various studies and led to controversial discussions.

The moral/conventional distinction was introduced to psychology by Turiel (1983). According to him, the purpose of conventional rules is to coordinate behavior in social systems. They gain their binding status by consensus within a given society, and they are arbitrary in that different agreements could have led to alternative conventions which would be just as feasible or appropriate. This implies that it is impossible to know whether a given action is in accordance with present social conventions by looking at the *action itself*. To know that it is appropriate to address your teacher by her last name, for example, you need to know that your society agreed that this is the proper way to behave since there is nothing intrinsically bad about addressing your teacher by her first name.

By contrast, the main distinguishing feature of moral rules, according to Turiel (1983), is that they are *not* arbitrary in the way conventions are. The reason for this is that they are concerned with certain *contents*: They regulate actions that have intrinsic consequences related to harm, fairness, or justice. A prototypical example of a moral transgression is that of a child pushing another child off a swing just because she wants to use it instead. According to Turiel, this act can be directly classified as harmful by any human observer familiar with pain regardless of the cultural context in which the act occurs.

Turiel (1983) argued that the moral and the conventional are separate domains of social knowledge, which are acquired and processed independently. In his view, moral rules can be empirically distinguished from conventional rules based on what Kelly, Stich, Haley, Eng, and Fessler (2007) called *signature moral pattern*. This term refers to a characteristic set of reactions people usually exhibit when they judge transgressions of prototypical moral rules: People supposedly consider moral rules, in contrast to conventional rules, to be valid even if they are not enforced by an authority (authority independence). Furthermore, moral rules are considered to be universally valid for all agents in similar circumstances across all times, places, and cultures (universality). Finally, people are expected to judge transgressions of moral rules as particularly serious and to justify their wrongness with reference to principles of harm, fairness, or justice.

An empirical paradigm to assess these signature patterns is the “moral/conventional task,” in which participants are presented with someone violating a prototypical moral rule (as in the swing example) or a social convention (as in the teacher example). If Turiel (1983) is right, then cases of moral transgression should reliably elicit the signature moral response pattern, while cases of conventional transgression should elicit the *signature conventional pattern* (i.e., judgments of authority dependence, lack of universality, decreased seriousness, and justification not related to harm, fairness, or justice; see Kelly et al., 2007). This pattern of response characteristics has been confirmed in a large number of empirical studies in diverse populations, including young children and people from different cultures (see Turiel, 2006, for an overview).

Despite its doubtless merit in advancing empirical research on moral psychology, the Turiel (1983) theory faces a number of conceptual and empirical problems. On the conceptual side, Turiel’s definition of morality seems to be at least in part a *petitio principii*: How people manage to recognize matters of fairness and justice seems to be just as much in need of explanation as how they recognize matters of morality. Intuitively, only the harm component seems to be a more basic concept, and more recent work trying to defend a content-based distinction mainly concentrates on harm (e.g., Royzman, Leeman, & Baron, 2009; Sousa, Holbrook, & Piazza, 2009).

The Turiel (1983) theory can also be contested on empirical grounds. In Turiel’s content-based approach to defining the moral domain, harm avoidance, fairness, and justice are tacitly assumed to be universal ends. We have already pointed out in the Introduction that cross-cultural research casts doubt on this universalist assumption.

Another empirical problem has been raised by Blair (1995). He questioned the assumption that only general cognitive capacities are needed to recognize moral transgressions. He showed that incarcerated psychopaths do not respond to moral transgressions with the signature moral pattern, even though they possess all the experiential and inferential capacities that according to Turiel (1983) are sufficient to distinguish the moral from the conventional. Blair’s (1995) proposal is that the human mind contains a specific module that is indispensable for this task, and which is selectively impaired in psychopathic individuals. He posits the existence of a violence inhibition mechanism (VIM) that is activated by perceptual key stimuli, mainly nonverbal

facial distress cues of suffering individuals. Once activated, the VIM triggers a withdrawal response in the observer, which he or she interprets as a moral emotion. According to Blair (1995), conventional transgressions lack these VIM specific distress cues. They therefore fail to activate the VIM and lead to a different response pattern.

It is notable that on Blair’s (1995) account, the main criterion for distinguishing between the moral and the conventional is shifted from a property of the *rules* (i.e., their content) to the activation of a cognitive structure in the *observer*. This focus has been adopted by other theorists, but there is considerable disagreement about the nature of this assumed cognitive structure. Nichols (2002), for example, doubts that a modular VIM as proposed by Blair (1995) would by itself be able to distinguish wrong from merely bad. Consider the example of a patient expressing agony while a nurse is changing the bandage of his wound (see Royzman et al., 2009). Although we instantly realize the patient’s suffering and even that the nurse is its proximal cause, we are not inclined to morally condemn her. It is not clear, however, why the observer’s VIM should not be activated in this example. It seems that we use some additional information when we make the moral/conventional distinction.

As already discussed in the section on “The Place of Emotions in Moral Judgments,” Nichols (2002, 2004) argues that this additional information is contained in a person’s “normative theory,” which is acquired in a cultural learning process. The normative theory contains all social rules, moral as well as conventional. Whether a given transgression elicits the signature moral pattern depends on whether the violated rule is backed by the activation of an emotion. This emotion need not be related to harm. For example, Nichols (2002) demonstrated that prototypically conventional transgressions can elicit the signature moral pattern if they are associated with disgust. His participants judged snorting loudly and spitting into cups at dinner tables to be universally and authority-independently wrong, especially those who scored high on a measure of disgust sensitivity (but see Royzman et al., 2009, for an alternative interpretation of the data).

Turiel’s (1983) distinction of two distinct social domains that can be empirically identified through unique signature patterns has also been questioned. Kelly et al. (2007) argue that the elements of the signature moral pattern on the one hand, and those of the signature conventional pattern on the other,

do not co-occur as monolithically as assumed by Turiel. For example, they point to the results of Haidt et al. (1993) and Nichols (2002) showing that people sometimes judge transgressions as universally wrong without justifying this assessment with notions of harm, fairness, or justice (see section on “Intuitionist Theories”). Thus, harm-related concerns do not seem *necessary* to elicit aspects of the signature moral pattern. Conversely, as Kelly et al. (2007) show, neither are they *sufficient* to do so. Instead of employing simple schoolyard transgressions of the kind Turiel (1983) and most of his followers used, Kelly and colleagues presented their participants with more complex harm-related cases from the adult world. For example, they asked their subjects whether it was okay for the captain of a modern U.S. cargo ship to whip one of his sailors as punishment for being drunk at work. Most participants responded that it was not. The researchers then went on to tell the same story, with the only difference that it was now supposed to have taken place several hundred years ago. The percentage of subjects considering the captain’s whipping behavior as wrong decreased significantly in this scenario. Similar changes were obtained when norm violations were fictitiously placed in faraway countries. Similarly, Kelly et al. (2007) showed that whether harmful actions are judged as wrong often depends on whether these actions were approved or forbidden by an authority. Taken together, these findings indicate that intrinsically harmful actions are not necessarily seen as universally or authority-independently wrong, contrary to what was assumed by Turiel (1983) and other harm-based approaches (Royzman et al., 2009; Sousa et al., 2009).

In light of this evidence, Sripada and Stich (2006) argue that the distinction between conventional and moral rules is not psychologically meaningful. At the same time, however, they share the intuition that not all social rules are treated identically. Rather than contrasting moral rules with conventional ones, they argue that there is a psychologically important subclass of rules that they call “norms.” Norms are not characterized by abstract philosophical principles or by specific contents, but mainly by the fact that people are *intrinsically motivated* to follow them.

In the Sripada and Stich (2006) model, people across all ages and cultures share the tendency to acquire and execute norms. The content of these norms, however, is assumed to be entirely determined by the social environment, and it need not (contrary to what Western philosophers have

assumed) be held to be universally valid. This view has the advantage that it simultaneously explains the difference between norms and conventions, while making no assumptions about the contents of the norms a specific culture selects. However, to date the mechanisms implementing this assumed norm acquisition device have only been sketched, so it is hard to see how it can be empirically tested against theories that assume an innate preparedness for the acquisition of specific moral rules. Also, it is not clear how the content-free norm view can predict which of the culturally endorsed rules will be assigned the status of norms as opposed to conventions. Moreover, given the lack of constraints on content a larger diversity of norms might be predicted than is actually observed. If, however, commonalities are explained as solutions to similar adaptive challenges all social groups face, which would be a plausible claim, then it may be implausible to exclude evolutionary processes from creating some of these commonalities (see Haidt & Joseph, 2007; Joyce, 2006).

Moral Dilemmas

The currently most discussed and studied moral dilemma in both philosophy and psychology is the trolley dilemma, which we already have encountered in previous sections. Trolley dilemmas have become the *Drosophila* for testing alternative philosophical and psychological theories of moral judgments in harm-based moral dilemmas. This dilemma is theoretically interesting for philosophers because it can be shown that people seem to reason according to consequentialist principles in some versions of the dilemma, but according to deontological rules in other versions (Foot, 1967; Kamm, 2007; Thomson, 1985; Unger, 1996). In the past decade a large number of psychological studies have been performed to pinpoint the factors underlying the different moral intuitions.

An influential study based on 5,000 subjects in 120 countries was performed by Hauser et al. (2007). This study is the primary evidence for moral grammar theory (see section on “Moral Grammar Theory”). Figure 19.1 illustrates two basic trolley cases used by Hauser and colleagues. In their variant of the *bystander* dilemma, the driver of a train heading toward five people on the track faints. Denise, a passenger, has the option to redirect the train toward a side track with one person. Eighty-five percent of the subjects responded “yes” to the test question that asked whether it is permissible

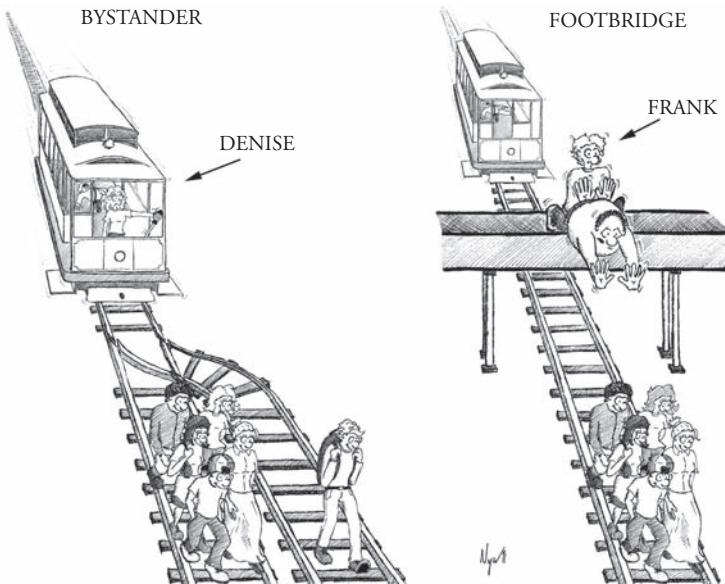


Fig. 19.1 Illustration of bystander and footbridge conditions as described by Hauser et al. (2007).

for Denise to turn the train. In the *footbridge* version of the dilemma, a runaway train is also heading toward five people. Here Frank stands on a bridge going over the tracks, realizing that he can stop the train with a heavy weight. The only available heavy weight is a large man standing next to him. Just twelve percent respond "yes" to the test question of whether it is permissible for Frank to shove the man. The effect was reliably observed in the studied countries, although subjects generally had difficulties justifying their intuitions (see also Cushman et al., 2006). The authors interpret the effect as evidence for the unconscious use of the *doctrine of double effect* (DDE, see section on "Moral Grammar Theory"), which allows harming a person as a side effect, but not as a means of saving more people.

However, the two conditions differ in a number of potentially relevant features, giving rise to various alternative explanations. First, in one condition the agent is on the threatening train, and therefore part of the danger, whereas in the other condition the agent is not part of the dangerous situation. This difference might contribute to the effect. Second, in the footbridge condition the agent could potentially sacrifice himself instead of the person next to him (although the reference to heaviness may suggest that the other person will stop the train more efficiently). At any rate, this is not an option in the bystander dilemma. Third, in one condition the act involves a morally irrelevant object (the switch), whereas in the other condition the act involves forceful contact

with the victim. Fourth, in one condition the intervention targets the threatening train, while in the other condition the victim is targeted. Fifth, the distance between the agent and the victim vary across conditions. Sixth, in one condition the potential alternative victim is only mentioned; in the other condition the victim is described as analogous to a heavy object. Seventh, the kind of death one imagines in the two conditions is more vivid and brutal in the footbridge than in the bystander condition; and eighth and last, the test questions differ in a way that can be expected to independently have an effect in the observed direction. Shoving a man is certainly considered less acceptable than turning a train even without the context of a trolley dilemma. A brief summary of the research of the past years is that it has been shown that almost all these confounding factors influence judgments, along with a number of others.

One plausible factor involves differences in *directness*. Previous research with other paradigms has shown that people find indirect harm less aversive than direct harm (Moore et al., 2008; Paharia et al., 2009; Royzman & Baron, 2002). Greene et al. (2009) have split this factor into three components: spatial proximity, physical contact, and personal force. Personal force refers to impacts on victims that are directly generated by muscular force of the agent. Touching and pushing a victim is an example of physical contact and personal force; using a pole for pushing would be an example of personal force without physical contact. Different versions of the

footbridge dilemma were compared in which the three components were pitted against each other. The results show that moral permissibility assessments were explained by the personal force factor. In additional studies it was shown that only personal force that is due to an intentional act is relevant.

Personal force does not provide a full account of intuitions in trolley dilemmas, however. Waldmann and his colleagues (Waldmann & Dieterich, 2007; Waldmann & Wiegmann, 2010) have constructed versions of the trolley dilemma in which the victims in all conditions were sitting in vehicles (which made their kind of death comparable), the agents were remote, the acts were equated, and neither physical contact nor personal force was necessary to act. For example, in the bystander variant a train heading toward a train with five passengers could be redirected to a side track in which a train with one passenger is located by pressing a button in a remote control center. In the contrasted footbridge analog, the setup is the same but now the train on the side track with one passenger could be redirected by pressing a button onto the main track, where this train would stop the threatening runaway train and thus save the five (Waldmann & Wiegmann, 2010). Participants reliably found the intervention in the first condition in which the threatening train was redirected ("threat intervention") more acceptable than the one in the second condition in which the train with the one victim was redirected ("victim intervention"), although no personal force was involved.

One possible explanation of the differences between threat and victim intervention is the DDE (e.g., Cushman et al., 2010; Mikhail, 2011; Royzman & Baron, 2002). Whereas in the threat intervention condition the victim is harmed as a side effect of saving the five, in the contrasted victim intervention condition the victim is used as a means to stop the runaway train. The doctrine of double effect can only be supported by indirect evidence, not by asking subjects. Cushman et al. (2006) have shown that this rule is not consciously accessible to subjects who are requested to provide a justification for their moral judgment. Other moral rules, however, such as the principle that touching and thereby harming a person (contact principle) is impermissible, can be consciously accessed.

A popular paradigm to test whether using a person as a means to save others is really particularly aversive is based on Thomson's (1985) loop idea. In this variant of the bystander dilemma, the side track loops back to the main track right before the

location where the five victims sit. This small variation turns the victim on the side track into a means to save the five. If the runaway train is redirected to the side track without being stopped by the person sitting there, it would go back to the main track and kill the five. Hauser et al. (2007) found that subjects judge the act in this condition more aversive than in the regular bystander (i.e., side effect) condition, but their experiment had the already mentioned confounds. Better-controlled studies did not find a difference (Greene et al., 2009; Waldmann & Dieterich, 2007; but see Sinnott-Armstrong, Mallon, McCoy, & Hull, 2008).

How else can the difference between threat and victim intervention be explained when personal force does not play a role in either condition? Waldmann and Wiegmann (2010) have proposed a *double causal contrast theory* to explain differences in intuitions in scenarios in which other relevant factors, such as distance, personal force, kind of victim, or kind of death have been held constant. The general idea motivating this theory is that people pay special attention to the intervention option when judging moral acceptability. Like all theories of moral judgments this theory predicts that reasoners are sensitive to the global contrast entailed by acting and nonacting (e.g., five victims vs. one), which explains why we differentiate between saving five or saving 1,000,000 (Bartels, 2008; Nichols & Mallon, 2006). However, whereas the DDE additionally is sensitive to the causal processes generated by the intervention (e.g., side effect vs. means), the double contrast theory assumes that we focus on the morally relevant target of intervention (i.e., threats or victims) and assess the harm directly caused by intervening on this target in contrast to the harm in which the target would be directly involved in the absence of the intervention. This local, counterfactual contrast focusing on the target of intervention will, according to this theory, heavily influence the acceptability rating.

How does the double causal contrast theory explain the two standard dilemmas? In the threat intervention condition, the proposed act can be summarized as redirecting the threat. Thus, the morally relevant target is the threatening trolley. To assess the local contrast, we need to focus on the direct harm caused by the target of intervention, which is one seriously harmed person. This outcome is contrasted with the direct harm caused by the target of intervention in the absence of the intervention, which in this condition are five

harmed people. In contrast, in the victim intervention condition, the proposed act can be described as redirecting the victim in the train on the side track. Thus, the local contrast will focus on the train with its single potential victim. Setting this train into motion will directly cause harm to this victim. The fact that five people are saved further in the future is an indirect, more remote consequence of the act and therefore not part of the local contrast. The proposed intervention is contrasted with what would happen to the target of intervention in the absence of an intervention. In this case the person sitting in the train on the safe track would remain unharmed. The local contrast implies that the act is harmful, which predicts the lowered acceptability ratings.

Double causal contrast theory explains why people are not only sensitive to how the victim is directly harmed by the intervention but also that they consider whether the victim would have been harmed in the absence of an intervention (Moore et al., 2008). Moreover, Waldmann and Wiegmann (2010) showed that people accept harming a victim as a means when the local contrast is favorable. In one of their experiments they described a trolley dilemma in which the runaway train threatening five carries a passenger. If nothing was done, this passenger would stay alive and the train would kill the five. The five can be saved, however, if an empty train is redirected toward the threatening train, derailing it by pushing its passenger, who would die in the process, against the emergency brakes. Although here the intervention directly kills one person who plays the role of a means to save the five, subjects find this act highly acceptable. They focus on the threatening train with its single victim as the target of intervention, which leads to a contrast between five and one dead person.

The debate about the role of the DDE focuses on causal and intentional factors underlying moral intuitions. However, there are many other factors influencing judgments in trolley dilemmas. Rai and Holyoak (2010) demonstrate that domain-general factors that have been identified in behavioral economics also affect judgments in trolley dilemmas. Subjects who were asked to generate many reasons in favor of the action paradoxically rated it as *less* permissible than those who generated fewer reasons. This is consistent with research by Schwarz (1998), who showed that ease of retrieval of justifications is used as indicator for the quality of an option in nonmoral consumer choice. Other factors are mood of subjects (Strohminger, Lewis, & Meyer, 2011; Valdesolo &

DeSteno, 2006), thinking styles (Bartels, 2008), preferred ethical position (Lombrozo, 2009), working memory (Moore et al., 2008), test question (see Kahane & Shackel, 2010), kind of victim (Uhlmann, Pizarro, Tannenbaum, & Ditto, 2009), vividness of death (Bartels, 2008), and the order of presenting different dilemmas (Iliev et al., 2009; Liao, Wiegmann, Alexander, & Vong, in press; Petrinovich & O'Neill, 1996; Wiegmann, Okan, & Nagel, in press). In short, it seems hopeless to look for the one and only explanation of moral intuitions in dilemmas. The research suggests that various moral and nonmoral factors interact in the generation of moral judgments about dilemmas.

The Role of Intention

A popular assumption implicit in many normative theories of morality is that we can only be held accountable for outcomes we have caused (Driver, 2008). We cannot possibly be responsible for bad events that are not directly or indirectly causally linked to our acts. In addition, most normative theories assign a special status to harm that was intentionally caused. For example, the doctrine of double effect (see sections on "Moral Grammar Theory" and "Moral Dilemmas") forbids intentionally harming a person, whereas unintentional harm may be permitted in some circumstances, even when the harmful outcome is foreseen. The interplay of intention and causal responsibility has also been central in descriptive theories of blame ascription (Alicke, 2000). A variety of recent studies have taken a closer look at the role of intentions and outcomes in moral judgments.

Cushman (2008) noted that different test questions may influence the relative contribution of these two components, intention and causation, in moral judgments. Cushman adopted a standard definition of intentional action according to which an act is intentionally performed if among other things the expected consequences are both desired and the act is believed to bring about these consequences (see Malle & Knobe, 1997). He used a story in which Jenny, the protagonist, is taking a course in sculpture and is assigned to work with a partner to weld together pieces of metal. The factors desire, belief, and consequences were manipulated independently using different cover stories: Jenny either desired or did not desire to burn her partner's hand, and she believed or did not believe that the act causes the harmful outcome. Moreover, it was varied whether the outcome did or did not occur. Cushman showed that judgments of blame and punishment are more

sensitive to whether the harmful outcome caused by the agent occurs, whereas judgments of wrongness and permissibility are more sensitive to the agent's belief with respect to harming someone. However, the belief factor was the strongest predictor for both kinds of judgments.

Although our normative intuitions imply that we only should be held responsible for outcomes that are under our causal control, a number of experiments about *moral luck* (Williams, 1982) have shown that negative outcomes that are not fully under the agent's control may also influence our judgments. A typical example of moral luck is the following scenario: A father who bathes his child in a tub answers the phone in the nearby living room after telling his child to stay put. He believes that his son will indeed stay put so that nothing bad will happen. The father is typically judged to be more morally blameworthy if his child drowns (an unlucky outcome) than if his child stays safe (a lucky outcome). Thus, in both cases the intentions and the knowledge are the same, but the outcomes vary due to unforeseen random factors. In psychology, the apparently inappropriate weight given to the outcome has been labeled *outcome bias*, which has been documented in many different scenarios (see Baron & Hershey, 1988; Gino et al., 2010).

One obvious theory explaining outcome biases is that people give undue weight to the valence of the outcome, even though the agent did not intend it and is not fully responsible for it. However, Young, Nichols, and Saxe (2010) proposed an alternative theory according to which moral luck depends strongly on belief attribution and only indirectly on the bad outcome. The theory claims that the bad outcome provides evidence that the unlucky agent's beliefs are erroneous. Holding an erroneous belief that can cause harm is blameworthy and therefore leads to harsher moral judgments. In contrast, in the condition of the lucky agent the outcome validates the correctness of the agent's prior beliefs.

Another recent controversy revolves around the causal relationship between intentions and moral judgments. Cushman (2008) and Young et al. (2010) adopt the traditional assumption that this relationship is unidirectional: The agent's intention determines the moral judgment of an act. However, consider the following example presented by Knobe (2003):

The vice-president of a company went to the chairman of the board and said, "We are thinking

of starting a new program. It will help us increase profits, but it will also harm the environment." The chairman of the board answered, "I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program." They started the new program. Sure enough, the environment was harmed. (p. 191)

In a second version of this scenario, the word "harm" was replaced by "help." When subjects were asked whether they think the chairman intentionally harmed the environment, eighty-two percent answered in the affirmative. In contrast, in the help condition seventy-seven percent said that the agent did *not* bring about the good side effect intentionally. Knobe concluded that in judging whether the side effect was brought about intentionally, the moral value of the side effect is crucial. People seem considerably more willing to say that a side effect was brought about intentionally when they regard it as bad than when they regard it as good. This finding suggests the opposite from what is traditionally assumed: The moral evaluation of the outcome seems to determine whether intentionality is attributed, not the other way around.

There have been a lot of attempts to explain the so-called side-effect effect (also known as Knobe effect). Most of the proposed explanations can broadly be put into two groups (see Feltz, 2007; Uttich & Lombrozo, 2010). One group explains the side-effect effect by claiming that moral evaluations actually play a role in our concept of intentional action. According to this group of theories, the concept of intentional action actually includes and can be determined by the moral value of the effects that are caused by the act (see, e.g., Knobe, 2010; Mele & Cushman, 2007; Nichols & Ulatowski, 2007).

The other group denies this claim and explains the side-effect effect by arguing that subjects' judgments of intentional action are biased. For instance, Adams and Steadman (2004) invoke conversational pragmatics and argue that people want to express blame for the agent in the negative side-effect condition by characterizing the outcome as intentional. Another supporter of this view is Nadelhoffer (2004), who claims that subjects' judgments are biased because they get emotionally affected by the bad side effect. Guglielmo and Malle (2010) believe that task demands forced participants to use the term "intentional:" When given a choice, most participants prefer to say that the agent brought about the bad side effect *knowingly* rather than *intentionally*.

(similar to what the DDE would predict; see section on “Moral Grammar Theory”).

Recently, Uttich and Lombrozo (2010) offered an interesting explanation that does not fall into either of these groups. They propose a rational explanation of the side-effect effect according to which the asymmetry in intentionality judgments arises because behavior that conforms to norms (moral or otherwise) is less informative regarding the mental state of the actor than norm-violating behavior (see Machery, 2008, for another theory of this kind). Prescriptive norms give us a positive reason to act, regardless of our intentions. Hence, a behavior that conforms to a norm does not tell us much about the agent’s intentions. In contrast, violating a norm provides us with positive evidence about the agent’s mental state. For example, in Knobe’s examples the norm is not to harm the environment. Thus, when the chairman starts a program that helps the environment, we cannot tell whether he intends this or just follows a norm, whereas in the contrast case the norm-violating behavior provides us with strong evidence of an intention to harm the environment. This theory is not restricted to moral norms; rather, it applies to all norms, moral or nonmoral. Uttich and Lombrozo (2010) could show that virtually the same asymmetry can be observed when the cover stories mention the conventional norm that specific cars usually have a dark color. Knobe (2010) has recently offered a similar explanation, but in contrast to Uttich and Lombrozo he highlights the role of *moral* norms.

In sum, the present research indicates that the role of intention is far more complex than previously thought. Intentions are unobservable states that need to be inferred. Apparently, a large number of factors, including observed behavior, outcomes, causal structure, rationality assumptions, and norms, contribute to these attributions. Moreover, our language allows for subtle differentiations between different types of intentionality (e.g., desire, want, intend, foresee), which form a complex network with other factors underlying moral judgment.

An interesting direction for future research might be to take a closer look at the role of intentions in different moral domains and in different cultures. As for domain differences, Young and Saxe (2011) have shown that intentions are assigned more weight for moral judgments of harm violations, like assault, compared to purity violations, like incest. Thus, differences in the role of intentions between different cultures may arise due to differences in

the culture-specific importance of moral domains. However, it is also possible that cultures differ within domains. In Western societies, the intentions of the agent are viewed as very important when assessing moral accountability, possibly more than in other cultures. Intentional transgressions of moral rules are typically condemned much more than accidental transgressions. In contrast, Rai and Fiske (2011) point out that in honor cultures, a woman who has sexual relations outside marriage, *even against her will*, defiles her family and is therefore punished.

Sacred/Protected Values

A characteristic feature of some moral values is that they resist trade-offs with other values. For example, many people find it impossible, inappropriate, or even outright abhorrent to put a price on human lives, friendships, democratic votes, or the preservation of the environment. It seems that some people ascribe infinite values to such entities, in that they would not accept any amount of any other good (especially not monetary ones) as compensation for the destruction or compromise of them. Such values have been termed “sacred” (Tetlock, Peterson, & Lerner, 1996) or “protected” (Baron & Spranca, 1997). Although both terms refer to the same phenomenon, the corresponding lines of research analyze it from different theoretical viewpoints, yielding different implications and even partially incompatible conclusions.

Tetlock and his colleagues describe sacred values (SVs) in their cultural context and analyze their social psychological functions. According to the revised Value Pluralism Model (Tetlock et al., 1996), people value different things for different reasons. When it comes to interpersonally relevant entities (such as intimate relationships, human rights, or religious symbols), people feel they have a commitment to others within their cultural community; they need to respect these entities in order to demonstrate that they are an estimable member of the community. The categorical nature of these commitments implies that within these sacred domains of social life, favors and goods are usually exchanged without numerical comparison (Fiske & Tetlock, 1997). If people compromise the respective values by trading them off against secular values (such as money, time, or convenience), they disqualify themselves from important social roles. Sacred and secular values are constitutively incommensurable; they cannot be sensibly compared, and mere attempts of comparison can destroy the SVs.

The motivation of people to hold SVs is to preserve their identity as full-fledged moral beings. If they witness others engaging in or merely contemplating taboo trade-offs, they typically react with moral outrage, a unitary response pattern consisting of harsh trait attributions, anger or contempt, and strong punitive impulses toward the offender (Tetlock, Kristel, Elson, Green, & Lerner, 2000). Due to unavoidable resource constraints in the real world, however, people are often forced to trade off SVs themselves. In such cases, they go to great lengths to conceal these trade-offs, for example, by means of decision avoidance or rhetorical obfuscation. Thus, people can be portrayed as both unapologetic defenders of SVs and at the same time as experts in finding ways to camouflage or overlook transgressions (Tetlock, 2003). Despite this discrepancy, Tetlock does not see people as hopeless hypocrites but instead as intuitive theologians striving to “[protect] sacred values from secular encroachments” (Tetlock, 2002, p. 452). Their rigidity is not seen as irrational but instead as serving important psychological functions and, on a larger scale, preventing subversion of meaningful cultural institutions (Fiske & Tetlock, 1997).

Baron approaches protected values (PVs) in the framework of the heuristics and biases program (see also section on “Moral Heuristics”). The main idea is that PVs are derived from deontological rules about acts (e.g., “do not kill”), irrespective of the consequences (Baron & Spranca, 1997). These rules are usually adaptive if treated as rules of thumb, but they may sometimes lead to suboptimal outcomes if they are unreflectively generalized to all contexts (Baron, 1998). In contrast to Tetlock, Baron reduces human values to a single utility metric, treating PVs as biases and stressing the problems they create for a utilitarian analysis.

One implication of the basis for PVs in absolute deontological rules is *quantity insensitivity*. For example, it seems to make only a small difference for people whether an act leads to greater or lesser harm to one of their PVs (Baron & Spranca, 1997), and some people seem to find it equally wrong to compromise a PV once or twice (Ritov & Baron, 1999). Another feature of deontological rules is that they usually prohibit harmful acts but not omissions, since prohibiting the latter would produce potentially unlimited obligations (Baron & Miller, 2000). Thus, PVs are seen as a source of omission bias (Baron & Ritov, 2009; Ritov & Baron, 1999) because actions are more likely to compromise PVs than omissions. For example, Ritov and Baron (1999) presented their

subjects with a scenario in which 20 species of fish living in a river would become extinct unless a dam was opened. However, opening the dam would cause the extinction of two different species living downstream which would otherwise survive. People with a PV against extinguishing species were especially unwilling to open the dam, even though this decision would result in a greater net amount of damage to their cherished natural resource.

Many people seem to readily endorse statements implying PVs when asked directly (e.g., “This should be prohibited no matter how great the benefits from allowing it;” Baron & Spranca, 1997, p. 7). However, according to Baron and Leshner (2000), such judgments may be the result of reflexive, incomplete thinking which can be overcome quite easily. When PVs are challenged with realistic counterexamples involving extremely high benefits or low probabilities for harm to PVs, many people relativize their absolute claims. This finding is taken to indicate that expressions of PVs should not be taken too seriously.

This remarkable tension between rigidity and flexibility has recently been interpreted differently by Bartels, Medin, and colleagues. Instead of seeing deontological judgments as an impediment for consequentialist judgments, they regard both as often positively correlated across people (Iliev et al., 2009). That is, people holding PVs that rigidly prohibit certain actions in one task can be shown to be especially sensitive to consequences of these actions in different tasks, compared to people without PVs. Whether they give more weight to means or ends is largely a function of their attentional focus, which in turn is crucially affected by domain-general individual thinking styles, as well as low-level features of the task, such as framing and context effects (Bartels, 2008). For example, Bartels and Medin (2007) argued that the framing of the response options used by Ritov and Baron (1999) in the river diversion scenario (“Would you open the dam? Yes/No,” followed by a measure for quantity sensitivity) directs the subjects’ attention to the act of killing species. In this condition many people maintain a categorical prohibition against this act, which leads them to express a PV. Bartels and Medin (2007) went on to show that reframing the response alternatives so that they deflect attention away from the action itself to the consequences (by having subjects choose from a list of alternatives the maximum number of species living downstream they would be willing to kill by opening the dam to save the twenty species at risk)

leads people with PVs to become *more* quantity sensitive, and less likely to show an omission bias than those without PVs. It seems as if the moral issue at stake is more central for people with PVs, and that they show amplified reactions in whatever direction their attention is steered by the task at hand (but see Baron & Ritov, 2009). In general, research on sacred and protected values provides an interesting test case showing that theories of moral judgments need to combine domain-specific cognitions (e.g., moral values) and domain-general mechanisms (e.g., attention).

Conclusions and Future Directions

The recent close cooperation both within psychology and across different disciplines has led to numerous new insights about morality. Summarizing the research from the viewpoint of a cognitive psychologist, three general research foci can be identified. First, many researchers have been interested in exploring the role of emotions and affects in moral judgments (see sections on “Emotion-based Theories” and “Dual-Process Theory”). This interest was initially motivated by a critique of previous paradigms (e.g., Kohlberg, 1981) in which conscious reasoning and rational deliberations were given a central place. In contrast, the more recent research has shown that many judgments are based on intuitions that are unconsciously elicited and are often accompanied by affects and emotions. The exact role of emotions is still not entirely clear. Emotions may precede or follow judgments, they may be constitutive for judgments, or they may be independent of rational judgment processes. A likely outcome of this debate may be that all of these possibilities occur, although we still need to know the boundary conditions of the different possibilities.

Second, the research on intuitions and emotions has largely addressed the global question of how reasoning and emotions in general are interrelated, but it has neglected the issue how specific intuitive judgments are caused. Thus, based on this research it is often impossible to make specific predictions about judgments for specific moral issues. The research has frequently been abstractly organized around a dichotomy between conscious reasoning and unconsciously elicited intuitions, which may have led to a neglect of research about the cognitive processes eliciting intuitions. In cognitive psychology, very few processes, not even logical reasoning and problem solving, are considered under full conscious control (see Evans, Chapter 8). Rather,

cognitive theories specify the often unconscious information processing steps leading from an eliciting situation to a judgment. Although we still know little about these processes, some researchers have made progress in recent years specifying moral rules (e.g., doctrine of double effect) or moral heuristics underlying the appraisal of moral scenarios.

Third, an overarching question motivating most research on moral judgment is whether moral cognitions are special, or whether they represent just specific contents that otherwise can be handled by domain-general theories. The present research suggests that there is no innate specialized module devoted to morality that is encapsulated from other cognitive processes. Many studies that were motivated by domain-general theories, for example, behavioral economics, judgment and decision-making theories, or attention theories, have shown that moral reasoning is not an isolated process but rather recruits domain-general processes that may lead to phenomena also found in other domains. On the other hand, a full reduction of moral cognitions to general cognitions also seems implausible. Moral judgments use moral rules, moral values, or norms that have characteristics that differ from the general class of rules. They are typically accompanied by strong affect and emotions, which endow them with a force that goes beyond general conventional norms. Moral rules or norms are typically viewed as authority independent, as ends that have to be honored, as particularly important, and by some people as universally valid. Thus, there is a consensus in the literature that humans are born with dispositions to honor norms that manifest themselves in moral judgments. Whether beyond the general capacity to acquire norms, there is also an innate capacity that predisposes humans to acquire specific *moral* rules, is an open question that is currently strongly debated.

In this review, we concentrated on research about explicit judgment tasks. Some researchers have questioned whether studying isolated judgments, especially with controlled experimental tasks, is ecologically valid (Gigerenzer, 2010). Our position is that we should not primarily study moral judgments to predict behavior, but rather to understand how people judge what is right or wrong. People’s opinions about moral issues, such as abortion, capital punishment, health, or food, are important factors shaping our society. However, it can be argued that implicit judgments are also reflected in *actions*. Although it is well known that moral judgments are not strongly correlated with corresponding actions, it

is interesting to compare explicit with more implicit moral evaluations. There are several interesting lines of research investigating actual behaviors that can be viewed as indicators of implicit moral judgment. For example, it has been shown that people paradoxically feel licensed to behave in morally dubious ways (e.g., cheating, lying, not donating to charity, or making uncooperative decisions) when they have activated a particularly positive view of their moral self (Mazar, Amir, & Ariely, 2008; Mazar & Zhong, 2010; Sachdeva, Iliev, & Medin, 2009). Conversely, they feel compelled to act particularly morally when their moral self-image is threatened ("moral cleansing behavior," see Sachdeva et al., 2009; Tetlock et al., 2000), demonstrating the importance of self-regulatory processes for implicit judgments underlying moral behavior.

People's implicit judgments concerning issues of fairness, altruism, cooperation, and punishment have also been assessed using behavioral measures. Fairness has been extensively investigated in simple bargaining games, primarily in the Ultimatum and Dictator Games, which investigate when subjects would reject unfair distributions of goods even when this implies that they would not get anything (e.g., Camerer, 2003; Camerer & Smith, Chapter 18). Common good games, which study individuals competing with other members of a group for common resources, have been used to obtain behavioral measures of cooperation, defection, and punishment (e.g., Fehr & Gächter, 2000). There is also a huge literature on altruism and prosocial behavior (see Batson, 2011).

We have seen that, although there seems to be an explosion of research on morality in recent years, many questions remain unanswered. Here we just list a few of these questions that seem particularly pressing from the viewpoint of cognitive psychology. For example, we know very little about the appraisal processes leading to moral judgments. Most of the theories dealing with appraisal have been developed in the context of very limited paradigms (e.g., trolley problems), so that the generality of these theories is unknown. Moreover, oversimplified theories of the representation of moral norms have postulated rules that seem to only superficially fit the investigated task. "Do no harm," for example, is certainly a rule that often seems plausible, but it does not capture the context sensitivity that people's judgments display. Thus, if a rule-based account is chosen, a much more complex system of rules needs to be specified, which includes boundary conditions and exceptions. Moreover, if research

on categorization is taken as a model (see Rips et al., Chapter 11), we need to ask whether rules are the only plausible format for the representation of moral knowledge or whether other representational devices, such as exemplars, prototypes, schemas, or analogies, also play a role.

We expect more research concerning the interplay of domain-general and domain-specific process in moral judgments. As the research on trolley dilemmas (see section on "Moral Dilemmas") shows, it seems necessary to negotiate the relative role of these processes for each target problem separately. There has been a tendency in the field to overstate findings as evidence for the use of grand philosophical positions. In our view, it seems implausible to argue that a sociopath reasons like a consequentialist when a much simpler account can be found. The fact that somebody finds smothering a baby abhorrent, or that somebody finds it preferable that one person instead of 1,000,000 people dies, does not turn this person into a deontologist or a consequentialist. It seems more plausible to pinpoint the reason for different judgments on more local factors, such as selective attention to specific aspects of a situation or deficits of affective processing.

Our review was largely limited to studies focusing on Western moral norms (e.g., prohibition of harm), which have been central in studies on the cognitive and affective foundations of moral judgment. The explanation for this one-sidedness is that both researchers and research subjects typically have a Western background (Henrich, Heine, & Norenzayan, 2010). Although anthropology has collected massive evidence showing that there is more to morality than concerns about harm or fairness/justice, most of this research so far is descriptive. We know that other cultures often endorse other norms, but we do not know how moral cognitions in other societies differ from ours. Do people in other cultures employ the same cognitive processes but invoke different moral rules, or are the cognitive processes underlying judgments different in other cultures? The most likely answer is that both possibilities may turn out to be true. If specified very abstractly, a process such as attentional focus will certainly influence judgments in different cultures, although the target of the focus will of course shift. On the other hand, we do not know whether general regularities that go beyond specific rules but are less abstract than attention universally play a similar role. For example, the section on "The Role of Intention" highlighted the role of intention in moral

blame. An interesting question might be whether intentional attributions and the weighing of intentions are similar in different domains and in different cultures. In sum, moral cognitions are most certainly an interesting topic for future research, but we have only started to understand this fascinating competency.

Acknowledgments

We thank K. Holyoak, T. Rai, and H. Rakoczy for helpful comments.

References

- Adams, F., & Steadman, A. (2004). Intentional action in ordinary language: Core concept or pragmatic understanding? *Analysis*, 64, 173–181.
- Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, 126, 556–574.
- Baron, J. (1994). Nonconsequentialist decisions. *Behavioral and Brain Sciences*, 17, 1–42.
- Baron, J. (1998). *Judgment misguided: Intuition and error in public decision making*. New York: Oxford University Press.
- Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, 54, 569–579.
- Baron, J., & Leshner, S. (2000). How serious are expressions of protected values? *Journal of Experimental Psychology: Applied*, 6, 183–194.
- Baron, J., & Miller, J. G. (2000). Limiting the scope of moral obligations to help: A cross-cultural investigation. *Journal of Cross-Cultural Psychology*, 31, 703–725.
- Baron, J., & Ritov, I. (2009). Protected values and omission bias as deontological judgments. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin (Eds.), *Moral judgment and decision making. The psychology of learning and motivation: Advances in research and theory* (pp. 133–167). San Diego, CA: Elsevier.
- Baron, J., & Spranca, M. (1997). Protected values. *Organizational Behavior and Human Decision Processes*, 70, 1–16.
- Bartels, D. M. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition*, 108, 381–417.
- Bartels, D. M., & Medin, D. L. (2007). Are morally motivated decision makers insensitive to the consequences of their choices? *Psychological Science*, 18, 24–28.
- Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition*, 121, 154–161.
- Batson, C. D. (2011). *Altruism in humans*. Oxford: Oxford University Press.
- Bennis, W. M., Medin, D. L., & Bartels, D. M. (2010). The costs and benefits of calculation and moral rules. *Perspectives on Psychological Science*, 5, 187–202.
- Bimrose, K. (2008). *Rational decisions*. Princeton, NJ: Princeton University Press.
- Blair, R. J. R. (1995). A cognitive developmental approach to morality: Investigating the psychopath. *Cognition*, 57, 1–29.
- Bucciarelli, M., Khemlani, S., & Johnson-Laird, P. N. (2008). The psychology of moral reasoning. *Judgment and Decision making*, 3, 121–139.
- Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton, NJ: Princeton University Press.
- Caruso, E. M., & Gino, F. (2011). Blind ethics: Closing one's eyes polarizes moral judgments and discourages dishonest behavior. *Cognition*, 118, 280–285.
- Chomsky, N. (1957). *Syntactic structures*. The Hague, Netherlands: Mouton.
- Ciaramelli, E., Muccioli, M., Làdavas, E., & di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive Affective Neuroscience*, 2, 84–92.
- Crain, W. C. (1985). *Theories of development*. Upper Saddle River, NJ: Prentice-Hall.
- Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, 108, 353–380.
- Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from non-moral psychological representations. *Cognitive Science*, 35, 1052–1075.
- Cushman, F., Young, L., & Greene, J. D. (2010). Multi-system moral psychology. In J. M. Doris & The Moral Psychology Research Group (Eds.), *The moral psychology handbook* (pp. 47–71). Oxford, England: Oxford University Press.
- Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, 17, 1082–1089.
- Damasio, A. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York: Putnam.
- Driver, J. (2008). Attributions of causation and moral responsibility. In W. Sinnott-Armstrong (Ed.), *Moral psychology. Vol. 2: The cognitive science of morality: Intuition and diversity* (pp. 423–440). Cambridge, MA: MIT Press.
- Dupoux, E., & Jacob, P. (2007). Universal moral grammar: A critical appraisal. *Trends in Cognitive Sciences*, 11, 373–378.
- Dwyer, S. (2006). How good is the linguistic analogy? In P. Carruthers, S. Lawrence, & S. Stich (Eds.), *The innate mind: Culture and cognition* (pp. 237–256). New York: Oxford University Press.
- Eskine, K. J., Kacinik, N. A., & Prinz, J. J. (2011). A bad taste in the mouth: Gustatory disgust influences moral judgment. *Psychological Science*, 22, 295–299.
- Evans, J. St. B. T. (2007). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 59, 255–278.
- Eyal, T., Liberman, N., & Trope, Y. (2008). Judging near and distant virtue and vice. *Journal of Experimental Social Psychology*, 44, 1204–1209.
- Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90, 980–994.
- Feltz, A. (2007). The Knobe effect: A brief overview. *Journal of Mind and Behavior*, 28, 265.
- Feltz, A., & Cokely, E. T. (2009). Do judgments about freedom and responsibility depend on who you are? Personality differences in intuitions about compatibilism and incompatibilism. *Consciousness and Cognition*, 18, 342–350.
- Fiske, A. P., & Tetlock, P. E. (1997). Taboo trade-offs: Reactions to transactions that transgress the spheres of justice. *Political Psychology*, 18, 255–297.
- Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, 5, 5–15.

- Gigerenzer, G. (2010). Moral satisficing: Rethinking moral behavior as bounded. *Topics in Cognitive Science*, 2, 528–554.
- Gigerenzer, G., Todd, P. M., & The ABC Research Group (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.
- Gilligan, C. (1982). *In a different voice: Psychological theory and women's development*. Cambridge, MA: Harvard University Press.
- Gino, F., Shu, L. L., & Bazerman, M. H. (2010). Nameless + harmless = blameless: When seemingly irrelevant factors influence judgment of (un)ethical behavior. *Organizational Behavior and Human Decision Processes*, 111, 93–101.
- Graham, J., Haidt, J., & Nosek, B. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96, 1029–1046.
- Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology. Vol. 2: The cognitive science of morality* (pp. 35–79). Cambridge, MA: MIT Press.
- Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, 111, 364–371.
- Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, 107, 1144–1154.
- Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44, 389–400.
- Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI study of emotional engagement in moral judgment. *Science*, 293, 2105–2108.
- Guglielmo, S., & Malle, B. F. (2010). Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Personality and Social Psychology Bulletin*, 36, 1635–1647.
- Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108, 814–834.
- Haidt, J. (2007). The new synthesis in moral psychology. *Science*, 316, 998–1002.
- Haidt, J., & Hersh, M. A. (2001). Sexual morality: The cultures and reasons of liberals and conservatives. *Journal of Applied Social Psychology*, 31, 191–221.
- Haidt, J., & Joseph, C. (2007). The moral mind: How 5 sets of innate moral intuitions guide the development of many culture-specific virtues, and perhaps even modules. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind* (Vol. 3, pp. 367–391). New York: Oxford University Press.
- Haidt, J., & Kesebir, S. (2010). Morality. In S. Fiske, D. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (5th ed., pp. 797–832). Hoboken, NJ: Wiley.
- Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology*, 65, 613–628.
- Harman, G., Mason, K., & Sinnott-Armstrong, W. (2010). Moral reasoning. In J. M. Doris & The Moral Psychology Research Group (Eds.), *The moral psychology handbook* (pp. 206–245). Oxford, England: Oxford University Press.
- Hauser, M. (2006). *Moral minds: How nature designed our universal sense of right and wrong*. New York: HarperCollins.
- Hauser, M. D., Cushman, F. A., Young, L., Jin, R., & Mikhail, J. M. (2007). A dissociation between moral judgment and justification. *Mind and Language*, 22, 1–21.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33, 61–83.
- Huebner, B., Dwyer, S., & Hauser, M. (2009). The role of emotion in moral psychology. *Trends in Cognitive Sciences*, 13, 1–6.
- Hume, D. (1960). *An enquiry concerning the principles of morals*. La Salle, IL: Open Court. (Original work published in 1777).
- Iliev, R., Sachdeva, S., Bartels, D. M., Joseph, C., Suzuki, S., & Medin, D. L. (2009). Attending to moral values. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin, (Eds.), *Moral judgment and decision making. The psychology of learning and motivation: Advances in research and theory* (pp. 169–192). San Diego, CA: Elsevier.
- Janoff-Bulman, R., Sheikh, S., & Hepp, S. (2009). Proscriptive versus prescriptive morality: Two faces of moral regulation. *Journal of Personality and Social Psychology*, 96, 521–537.
- Joyce, R. (2006). *The evolution of morality*. Cambridge, MA: MIT Press.
- Kahane, G., & Shackel, N. (2008). Do abnormal responses show utilitarian bias? *Nature*, 452, 5.
- Kahane, G., & Shackel, N. (2010). Methodological issues in the neuroscience of moral judgment. *Mind and Language*, 25, 561–582.
- Kahane, G., Wiech, K., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (in press). The neural basis of intuitive and counterintuitive moral judgment. *Social Cognitive and Affective Neuroscience*.
- Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In K. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 267–294). Cambridge, England: Cambridge University Press.
- Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.
- Kamm, F. M. (2007). *Intricate ethics*. Oxford, England: Oxford University Press.
- Kant, I. (1959). *Foundation of the metaphysics of morals* (L. W. Beck, Trans.). Indianapolis, IN: Bobbs-Merrill. (Original work published in 1785).
- Kelly, D., Stich, S., Haley, K. J., Eng, S. J., & Fessler, D. M. T. (2007). Harm, affect, and the moral/conventional distinction. *Mind and Language*, 22, 117–131.
- Kern, M. C., & Chugh, D. (2009). Bounded ethicality: The perils of loss framing. *Psychological Science*, 20, 378–384.
- Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, 63, 190–194.
- Knobe, J. (2006). The concept of intentional action: A case study in uses of folk psychology. *Philosophical Studies*, 130, 203–231.
- Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, 33, 315–329.
- Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., & Hauser, M. (2007). Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature*, 446, 908–911.
- Kohlberg, L. (1981). *The philosophy of moral development*. San Francisco, CA: Harper.
- Laham, S. M., Alter, A. L., & Goodwin, G. P. (2009). Easy on the mind, easy on the wrongdoer: Discrepantly fluent

- violations are deemed less morally wrong. *Cognition*, 112, 462–466.
- Lazarus, R. S. (1991). Cognition and motivation in emotion. *American Psychologist*, 46, 352–367.
- Liao, S. M., Wiegmann, A., Alexander, J., & Vong, G. (in press). Putting the trolley in order: Experimental philosophy and the loop case. *Philosophical Psychology*.
- Litman, L., & Reber, A. S. (2005). Implicit cognition and thought. In K. J. Holyoak & R. G. Morrison (Eds.), *Cambridge handbook of thinking and reasoning* (pp. 431–453). New York: Cambridge University Press.
- Lombrozo, T. (2009). The role of moral commitments in moral judgment. *Cognitive Science*, 33, 273–286.
- Machery, E. (2008). The folk concept of intentional action: Philosophical and experimental issues. *Mind and Language*, 23, 165.
- Machery, E., & Mallon, M. (2010). The evolution of morality. In J. M. Doris & The Moral Psychology Research Group (Eds.), *The moral psychology handbook* (pp. 3–46). Oxford, England: Oxford University Press.
- Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, 33, 101–121.
- Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, 45, 633–644.
- Mazar, N., & Zhong, C. (2010). Do green products make us better people? *Psychological Science*, 21, 494–498.
- McGuire, J., Langdon, R., Coltheart, M., & Mackenzie, C. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. *Journal of Experimental Social Psychology*, 45, 577–580.
- Mele, A., & Cushman, F. (2007). Intentional action, folk judgments, and stories: Sorting things out. *Midwest Studies in Philosophy*, 31, 184–201.
- Mendez, M. F., Anderson, E., & Shapira, J. S. (2005). An investigation of moral judgment in frontotemporal dementia. *Cognitive and Behavioral Neurology*, 18, 193–197.
- Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, 11, 143–152.
- Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. New York: Cambridge University Press.
- Moll, J., & de Oliveira-Souza, R. (2007). Moral judgments, emotions, and the utilitarian brain. *Trends in Cognitive Sciences*, 11, 319–321.
- Moll, J., de Oliveira-Souza, R., & Eslinger, P. J. (2003). Morals and the human brain: A working model. *NeuroReport*, 14, 299–305.
- Moll, J., de Oliveira-Souza, R., Eslinger, P. J., Bramati, I. E., Mourao-Miranda, J., Andreiuolo, P. A., & Pessoa, L. (2002). The neural correlates of moral sensitivity: A functional magnetic resonance imaging investigation of basic and moral emotions. *Journal of Neuroscience*, 22, 2370–2736.
- Moore, A. B., Clark, B. A., & Kane, M. J. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science*, 19, 549–557.
- Moore, A. B., Stevens, J., & Conway, A. R. (2011). Individual differences in sensitivity to reward and punishment predict moral judgment. *Personality and Individual Differences*, 50, 621–625.
- Nadelhoffer, T. (2004). On praise, side effects, and folk ascriptions of intentionality. *Journal of Theoretical and Philosophical Psychology*, 24, 196.
- Nichols, S. (2002). Norms with feeling: Towards a psychological account of moral judgment. *Cognition*, 84, 221–236.
- Nichols, S. (2004). *Sentimental rules: On the natural foundations of moral judgment*. New York: Oxford University Press.
- Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, 100, 530–542.
- Nichols, S., & Ulatowski, J. (2007). Intuitions and individual differences: The Knobe effect revisited. *Mind and Language*, 22, 346–365.
- Paharia, N., Kassam, K. S., Greene, J. D., & Bazerman, M. H. (2009). Dirty work, clean hands: The moral psychology of indirect agency. *Organizational Behavior and Human Decision Processes*, 109, 134–141.
- Parfit, D. (2011). *On what matters*. Oxford, England: Oxford University Press.
- Paxton, J. M., & Greene, J. D. (2010). Moral reasoning: Hints and allegations. *Topics in Cognitive Science*, 2, 511–527.
- Petrinovich, L., & O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology*, 17, 145–171.
- Piaget, J. (1932). *The moral judgment of the child*. London: Kegan Paul, Trench, Trubner and Co.
- Prinz, J. J. (2007). *The emotional construction of morals*. Oxford, England: Oxford University Press.
- Prinz, J. J. (2008). Acquired moral truths. *Philosophy and Phenomenological Research*, 77, 219–227.
- Prinz, J. J., & Nichols, S. (2010). Moral emotions. In J. M. Doris & The Moral Psychology Research Group (Eds.), *The moral psychology handbook* (pp. 111–146). Oxford, England: Oxford University Press.
- Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review*, 118, 57–75.
- Rai, T. S., & Holyoak, K. J. (2010). Moral principles or consumer preferences? Alternative framings of the trolley problem. *Cognitive Science*, 34, 311–321.
- Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Harvard University Press.
- Reyna, V. F., & Casillas, W. (2009). Development and dual processes in moral reasoning: A fuzzy-trace theory approach. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin (Eds.), *Moral judgment and decision making. The psychology of learning and motivation: Advances in research and theory* (pp. 207–236). San Diego, CA: Elsevier.
- Ritov, I., & Baron, J. (1999). Protected values and omission bias. *Organizational Behavior and Human Decision Processes*, 79, 79–94.
- Roedder, E., & Harman, G. (2010). Linguistics and moral theory. In J. M. Doris & The Moral Psychology Research Group (Eds.), *The moral psychology handbook* (pp. 273–296). Oxford, England: Oxford University Press.
- Royzman, E. B., & Baron, J. (2002). The preference for indirect harm. *Social Justice Research*, 15, 165–184.
- Royzman, E. B., Leeman, R. F., & Baron, J. (2009). Unsentimental ethics: Towards a content-specific account of the moral–conventional distinction. *Cognition*, 112, 159–174.
- Sachdeva, S., Iliev, R., & Medin, D. L. (2009). Sinning saints and saintly sinners: The paradox of moral self-regulation. *Psychological Science*, 20, 523–528.
- Sachdeva, S., & Medin, D. L. (2008). Is it more wrong to care less? The effects of “more” and “less” on the quantity (in) sensitivity of protected values. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual*

- Conference of the Cognitive Science Society* (pp. 1239–1243). Austin, TX: Cognitive Science Society.
- Scanlon, T. M. (1999). *What we owe to each other*. Cambridge, MA: Harvard University Press.
- Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin*, 34, 1096–1109.
- Schwarz, N. (1998). Accessible content and accessibility experiences: The interplay of declarative and experiential information in judgment. *Personality and Social Psychology Review*, 2, 87–99.
- Shweder, R. A., Much, N. C., Mahapatra, M., & Park, L. (1997). The “big three” of morality (autonomy, community, divinity) and the “big three” explanations of suffering. In A. Brandt & P. Rozin (Eds.), *Morality and health* (pp. 119–169). New York: Routledge.
- Simpson, E. (1974). Moral development research: A case study of scientific cultural bias. *Human Development*, 17, 81–106.
- Singer, P. (1979). *Practical ethics*. Cambridge, England: Cambridge University Press.
- Sinnott-Armstrong, W. (2008). Framing moral intuitions. In W. Sinnott-Armstrong (Ed.), *Moral psychology. Vol. 2: The cognitive science of morality: Intuition and diversity* (pp. 47–76). Cambridge, MA: MIT Press.
- Sinnott-Armstrong, W., Mallon, R., McCoy, T., & Hull, J. G. (2008). Intention, temporal order, and moral judgments. *Mind and Language*, 23, 90–106.
- Sinnott-Armstrong, W., Young, L., & Cushman, F. (2010). Moral intuitions. In J. M. Doris & The Moral Psychology Research Group (Eds.), *The moral psychology handbook* (pp. 246–272). Oxford, England: Oxford University Press.
- Sloman, S. A., Fernbach, P. M., & Ewing, S. (2009). Causal models: The representational infrastructure for moral judgment. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin (Eds.), *Moral judgment and decision making. The psychology of learning and motivation: Advances in research and theory* (pp. 1–26). San Diego, CA: Elsevier.
- Sousa, P., Holbrook, C., & Piazza, J. (2009). The morality of harm. *Cognition*, 113, 80–92.
- Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology*, 27, 76–105.
- Sripada, C. S. & Stich, S. (2006). A framework for the psychology of norms. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind* (Vol. 2, pp. 280–301). New York: Oxford University Press.
- Strohminger, N., Lewis, R. L., & Meyer, D. E. (2011). Divergent effects of different positive emotions on moral judgment. *Cognition*, 119, 295–300.
- Sunstein, C. (2005). Moral heuristics. *Behavioral and Brain Sciences*, 28, 531–573.
- Suter, R. S., & Hertwig, R. (2011). Time and moral judgment. *Cognition*, 119, 454–458.
- Tetlock, P. E. (2002). Social functionalist frameworks for judgment and choice: Intuitive politicians, theologians, and prosecutors. *Psychological Review*, 109, 451–471.
- Tetlock, P. E. (2003). Thinking the unthinkable: Sacred values and taboo cognitions. *Trends in Cognitive Sciences*, 7, 320–324.
- Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, 78, 853–870.
- Tetlock, P. E., Peterson, R. S., & Lerner, J. S. (1996). Revising the value pluralism model: Incorporating social content and context postulates. In C. Seligman, J. Olson, & M. Zanna (Eds.), *Ontario Symposium on Social and Personality Psychology: Values* (pp. 25–51). Mahwah, NJ: Erlbaum.
- Thomson, J. J. (1985). The trolley problem. *Yale Law Journal*, 94, 1395–1415.
- Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge, England: Cambridge University Press.
- Turiel, E. (2006). The development of morality. In N. Eisenberg, W. Damon, & R. M. Lerner (Eds.), *Handbook of child psychology. Vol. 3: Social, emotional, and personality development* (pp. 789–857). Hoboken, NJ: Wiley.
- Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., & Ditto, P. H. (2009). The motivated use of moral principles. *Judgment and Decision Making*, 4, 479–491.
- Unger, P. (1996). *Living high and letting die: Our illusion of innocence*. New York: Oxford University Press.
- Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, 116, 87–100.
- Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science*, 17, 476–477.
- Waldmann, M. R., & Dieterich, J. H. (2007). Throwing a bomb on a person versus throwing a person on a bomb: Intervention myopia in moral intuitions. *Psychological Science*, 18, 247–253.
- Waldmann, M. R., & Wiegmann, A. (2010). A double causal contrast theory of moral intuitions in trolley dilemmas. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2589–2594). Austin, TX: Cognitive Science Society.
- Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science*, 16, 780–784.
- Wiegmann, A., Okan, Y., & Nagel, J. (in press). Order effects in moral judgment. *Philosophical Psychology*.
- Williams, B. (1982). Moral luck. In *Moral luck. Philosophical papers 1973–1980* (pp. 20–39). Cambridge, England: Cambridge University Press.
- Williams, B. (1985). *Ethics and the limits of philosophy*. London: Fontana.
- Woodward, P. A. (Ed.) (2001). *The doctrine of double effect*. Notre Dame, IN: Notre Dame University Press.
- Wright, J. C., & Baril, G. (2011). The role of cognitive resources in determining our moral intuitions: Are we all liberals at heart? *Journal of Experimental Social Psychology*, 47, 1007–1012.
- Young, L., Nichols, S., & Saxe, R. (2010). Investigating the neural and cognitive basis of moral luck: It's not what you do but what you know. *Review of Philosophy and Psychology*, 1, 333–349.
- Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition*, 120, 149–298.

Motivated Thinking

Daniel C. Molden and E. Tory Higgins

Abstract

Once controversial, the idea that people's motivations can influence their cognitions now plays an important role in current research on thinking and reasoning. This chapter describes the effects on cognition of motivations that originate from three separate sources: (a) specific desired conclusions (e.g., perceptions of oneself as successful, loved, or in control); (b) more general desired conclusions (e.g., judgments that are as concise and unambiguous, or as accurate as possible); and (c) preferences for reaching such conclusions using particular types of judgment strategies (e.g., a focus on pursuing opportunities for gain versus protecting against the possibility of loss). Evidence is reviewed for the influence of each of these motivations on a variety of cognitive processes, illustrating that, in addition to being "cognitive misers" whose biases result from limited cognitive-processing capacity, people are "motivated tacticians" whose biases result from preferences for processing information in ways that serve their current motivational concerns.

Key Words: self-enhancement motivations, need for belonging, need for control, accuracy motivations, need for closure, regulatory focus, regulatory mode, regulatory fit

At one time or another, every one of us has engaged in "wishful thinking" or "let our hearts run away with our heads." That is, every one of us has experienced the effects of our motivations on our thoughts and behaviors. Over the past 10–15 years, researchers have shown a renewed interest in these types of experiences and conducted many investigations of how people's drives, needs, desires, motives, and goals can profoundly influence their judgment and reasoning across many different domains. This chapter provides an overview of such research on motivated thinking and describes how, in addition to being "cognitive misers" whose biases result from generally limited cognitive-processing capacity, people are "motivated tacticians" whose biases result from specific preferences for processing information in ways that serve their current motivational concerns.

Psychologists have long placed the impact of motivation on basic cognitive processes at the center of their analysis of human behavior (e.g., Allport, 1920; Lewin, 1935; Murray, 1938). However, for a period of time, this impact was questioned and effects related to motivated thinking were recast as purely cognitive phenomena stemming from imperfect information processing by imperfect perceivers (e.g., Bem, 1967; Nisbett & Ross, 1980). Yet once it became clear that debates about the workings of motivational *versus* cognitive processes in judgment and behavior were fruitless and counterproductive (e.g., Tetlock & Levi, 1982), a new wave of research emerged that emphasized the interface between these processes and began to identify more precisely the distinct mechanisms by which drives, needs, motives, and goals alter people's judgment and

actions (see, Dunning, 1999; Higgins & Molden, 2003; Kruglanski, 1996; Kunda, 1990).

In this chapter, we review this “second generation” of research on motivated thinking and consider some of the basic principles that have emerged. We examine two general classes of motivational influences that have been identified. The first involves people’s desires for reaching particular types of judgment *outcomes*, and the second involves people’s desires to use particular types of judgment *strategies*. Because we adopt a broad focus and discuss many different varieties of motivated thinking, we are selective rather than comprehensive in the phenomena described. The programs of research we highlight are those we believe to be representative of the larger literature and especially relevant not only to the study of reasoning but to other areas of experimental psychology.¹ After reviewing the separate influences of outcome- and strategy-based motivations, we conclude by suggesting potential directions for future research, giving special attention to possible interactions between a variety of motivations that might be simultaneously activated.

Outcome-Motivated Thinking

The most prominent approach to motivated reasoning, in both classic and contemporary perspectives, has been to examine the influence on people’s thought processes of their needs, preferences, and goals to reach desired outcomes (or avoid undesired outcomes). Although the types of desired outcomes studied have been highly diverse, they can be divided into two general classes: *directional* outcomes and *nondirectional* outcomes (see Kruglanski, 1996; Kunda, 1990). Individuals who are motivated by directional outcomes are interested in reaching specific desired conclusions, such as impressions of themselves as intelligent, caring, and worthy people (e.g., Dunning, 1999; Kunda, 1990); perceptions of belonging and social connection with others (e.g., Baumeister & Leary, 1995; Molden & Maner, in press); and feelings of control and understanding (e.g., Heine, Proulx, & Vohs, 2006; Pittman & D’Agostino, 1989). That is, such directional outcomes inherently specify the precise content of the conclusions people are motivated to reach. In contrast, individuals who are motivated by nondirectional outcomes have more general concerns such as reaching the most accurate conclusion possible (e.g., Fiske & Neuberg, 1990) or reaching a conclusion that is concise and unambiguous (e.g., Kruglanski

& Webster, 1996), whatever this conclusion may be. That is, nondirectional outcomes do not directly specify the precise content of the conclusions people are motivated to reach (although, as discussed in detail later, these outcomes can certainly still indirectly alter what conclusions they finally do reach).

Whether outcome motivation is directional or nondirectional, this motivation has been conceptualized as affecting thought and reasoning in the same general way: by directing people’s cognitive processes (e.g., their attributions, information search, or recall) in a manner that helps ensure that they reach the desired outcomes. In the next sections, we review several programs of research that have examined more closely the specific mechanisms by which this can occur, first in relation to motivations for directional outcomes and then in relation to motivations for nondirectional outcomes. Following this, we discuss several limitations of the effects of outcome motivation on reasoning and identify circumstances in which these motivations are most likely to have an impact.

Influences of Motivations for Directional Outcomes

Overall, the kinds of phenomena that have been studied most extensively in research on motivated thinking involve preferences for directional outcomes, that is, individuals’ desires to reach specific conclusions about themselves and others (see Dunning, 1999; Kunda, 1990; Murray, 1999). Overall, the outcomes that have, by far, received the most attention involve people’s well-documented preference for viewing themselves, and those close to them, in a generally positive manner (i.e., desires to preserve self-esteem; see Baumeister, 1998). However, more recently, other outcomes, such as desires to feel connected to others (Baumeister & Leary, 1995; Molden & Maner, in press) and desires for understanding of and control over one’s environment (Heine et al., 2006; Pittman & D’Agostino, 1989; Seligman, 1976) have also received more attention as well. Given the extensive reviews of esteem-related effects on motivated thinking available elsewhere (Kruglanski, 1996; Kunda, 1990; Molden & Higgins, 2005), we give priority to detailing the influence of other types of directional motivations here.

In the next sections, we review several effects of directional motivations involving many different cognitive processes, including attribution, evaluation of truth and legitimacy, attention and information

search, recall and knowledge activation, and the organization of concepts in memory. The studies described provide evidence for motivational effects on cognition using several different strategies. One is to correlate individual differences in the general importance of particular needs or goals with different judgment outcomes; another is to experimentally vary the contextual salience of particular motivations or the relevance of these motivations to the judgment at hand and examine the consequences; still another is to threaten or challenge people's progress toward fulfilling particular motivations and assess how people think and behave in response to this threat. Despite these variations in method, the basic logic of all of these studies is that greater motivational activation, whether from chronic personality orientations, temporary salience, or in response to a current challenge, should produce more motivationally consistent thinking and reasoning.

EFFECTS ON ATTRIBUTION

One early demonstration of the effects of motivation on judgment came from work on the explanation of behavior and the attribution of particular traits to oneself or others (see Kelley, 1973). Whereas many initial studies focused on how esteem-related motivations influenced people's claims of personal responsibility for their successes, and denial of responsibility for their failures (for a meta-analysis see Mezulis, Abramson, Hyde, & Hankin, 2004), recent programs of research have examined the attributional effects of motivations for belonging and feelings of control as well.

For example, one series of studies on how motives for social connection affect self-perception showed that students who viewed personality profiles of attractive, opposite-sex individuals as part of a "matchmaking" exercise later attributed the traits listed in the profiles more strongly to themselves, enhancing what they had in common with these individuals (Slotter & Gardner, 2009). Furthermore, such attributions emerged even for traits that students had identified weeks earlier as not describing them very well. Consistent with the motivational nature of this phenomenon, these effects were strongest among individuals who wanted to initiate a romantic relationship. Moreover, students who viewed identical personality profiles as part of a student-government campaign did not strongly attribute such traits to themselves.

Additional studies have demonstrated similar effects of motivations for social connection on

perceptions of others. Students who had learned that no one wanted to work with them on a group task later attributed more friendliness and likability to individuals they viewed in photographs than did students who had learned that many people wanted to work with them (Maner, DeWall, Baumeister, & Schaller, 2007). As is again consistent with the motivational nature of these effects, such findings were limited to individuals who were not overly fearful of negative evaluations from others, and thus sufficiently interested in pursuing *reconnection* following social exclusion (see also Maner et al., 2005).

Research on how motivations for feelings of control influence attribution has demonstrated analogous effects. Although some initial findings suggested that decreased feelings of control generally increase the effort and attention people give to their attributions overall (Pittman & D'Agostino, 1989), recent studies have indicated that people who experience less control selectively choose attributions that specifically bolster either their sense of personal control or perceptions that the world generally is an understandable and predictable place. For example, when people's feelings of control are threatened by reminding them that they will eventually die, they explain life-threatening events that happen to others more strongly in terms of these other individuals' unique traits and behaviors (Hirschberger, 2006). Such attributions presumably allow people to see the events as more avoidable and help them restore some sense of control (see also Fritzsche, Jonas, & Fankhänel, 2008). Similarly, when people are asked to relive instances in which they themselves were not in control, they subsequently attribute greater control to external institutions and agents, such as governments and God, than people who relive instances in which they had personal control (Kay, Gaucher, Napier, Callan, & Laurin, 2008). However, this effect occurs only when such external agents are perceived to be fair and benevolent and actually allow people to maintain a general sense of predictability and control (see also Kraus, Piff, & Keltner, 2009). Finally, when feelings of control are threatened through interactions with unpredictable mechanical objects or computer programs, people form stronger attributions of human-like agency and intention (Waytz et al., 2010), which also helps restore some feelings of understanding and predictability. Therefore, overall, increased motivations for both social connection and control produce attributions that better allow people to reach their desired conclusions.

EFFECTS ON EVALUATION OF TRUTH AND LEGITIMACY

In addition to favoring motivationally relevant attributions, people also preferentially evaluate the truth and legitimacy of information that supports (versus contradicts) motivationally relevant outcomes. A classic example of this effect, to which many researchers can perhaps relate, is that individuals reading about scientific studies that appear to support their cherished attitudes describe the studies as being better conducted, and their conclusions as being more valid, than individuals for whom the same studies are in conflict with their cherished attitudes (Lord, Ross, & Lepper, 1979; see also Beauregard & Dunning, 1998; Ditto, Scepansky, Munro, Apanovitch, & Lockhart, 1998; Kunda, 1987; Tetlock, 1998).

Similar phenomena have been observed for judgments about the rightness or wrongness of institutional practices. For example, people who were induced to believe that the institutions of the society to which they belonged (e.g., governments, universities, businesses) had greater control over their personal outcomes were more likely to judge the current institutional practices (whatever they were) as legitimate and right (Kay et al., 2009; see also Kay, Jost, & Young, 2005). Moreover, a nationally representative survey in the United States showed that people with stronger convictions concerning the moral wrongness of physician-assisted suicide not only viewed a 2006 decision by the Supreme Court supporting the legality of such measures as more unfair and personally unacceptable, they also questioned the legitimacy of the court itself and were more supportive of measures to reduce the court's ability to make these types of decisions. In contrast, people with stronger convictions concerning the moral rightness of physician-assisted suicide elevated the legitimacy of the court and its powers in such matters (Skitka, Bauman, & Lytle, 2009). Thus, overall, people give greater legitimacy to information and actions that are consistent with their desired conclusions but are more skeptical of the exact same information and actions when they are inconsistent with these conclusions.

EFFECTS ON ATTENTION AND INFORMATION SEARCH

Beyond affecting judgments of the validity of particular information, people's motivations can also influence whether they even attend to or notice motivationally consistent or inconsistent

information. Decades of research has investigated whether the information that people seek out or avoid is affected by their predominant attitudes, beliefs, and behaviors. A recent meta-analysis of this research revealed a sizable bias toward selectively attending to motivationally consistent information (Hart et al., 2009). Moreover, consistent with a motivational perspective of this effect, the bias was larger the more committed people were to their attitudes, beliefs, or behaviors and the more central the focal issue was to people's personal values.

Recent research has also demonstrated that such effects occur not only for conscious, explicit decisions about what types of information to process further but also for more unconscious and implicit attentional processes. For example, studies using eye-tracking and visual reaction-time measures have shown that (a) when motivations for belonging are threatened, people selectively fixate their attention on smiling faces that could signal opportunities for social inclusion (and not just on positive stimuli in general; DeWall, Maner, & Rouby, 2009); (b) when romantic motivations are activated (i.e., people are sexually aroused), they selectively fixate their attention on physically attractive faces of the opposite sex that could signal opportunities to fulfill those motivations (but not attractive faces of the same sex or unattractive faces of the opposite sex that do not signal such opportunities; Maner, Gailliot, Rouby, & Miller, 2007); and (c) when jealousy motivations are activated (i.e., people are concerned about romantic infidelity), they selectively fixate their attention on attractive faces of the same sex that could potentially represent romantic rivals (but not attractive faces of the opposite sex or unattractive faces of the same sex that do not represent rivals; Maner, Miller, Rouby, & Gailliot, 2009). Similarly, when people's motivations for control are threatened by the recall of a personal experience of being out of control, they subsequently attempt to impose order on their visual field and are more likely to report seeing coherent images in slides of white noise (Whitson & Galinsky, 2008; see also Proulx & Heine, 2009). Finally, when a particular stimulus has temporarily assumed greater value because it signals task completion or represents some kind of reward, people show sustained attention to this stimulus despite distractors (Raymond & O'Brien, 2009) and overestimate how close the stimulus appears to be (Balceris & Dunning, 2007, 2010).

The motivational influences that we have discussed in this section thus far all concern the *quality*

of attention and information search. However, such influences can also at times affect the *quantity* of information processing (Kruglanski, 1996). Evidence that is motivationally consistent receives decreased processing and quick acceptance, whereas evidence that is motivationally inconsistent receives more thorough processing and hesitant acceptance. That is, when people encounter information inconsistent with desired conclusions, they spend more time considering the validity of this evidence and exert more effort generating hypotheses about why it might be unreliable (see Ditto et al., 1998; Giner-Sorolla & Chaiken, 1997). For example, individuals high in prejudice show increased processing of information about behaviors of African Americans that contradicts, rather than confirms, negative stereotypes in attempts to explain why this counter-stereotypical behavior is simply a fluke (Sherman, Stroessner, Conrey, & Azam, 2005). Moreover, confirming that it is the quantity of processing that plays a critical role, when additional processing toward motivationally inconsistent information is inhibited by placing people under *cognitive load*, the typical motivational differences in people's judgments disappear (see also Plaks & Stecher, 2007; Skitka, Mullen, Griffin, Hutchinson, & Chamberlin, 2002).

EFFECTS ON RECALL AND KNOWLEDGE ACTIVATION

Another means by which motivations influence cognitive processing during judgment involves the activation and retrieval of knowledge stored in memory. Initial studies of esteem-related motivations demonstrated that such influences included (a) selective recall of motivationally consistent information from memory (e.g., Santioso, Kunda, & Fong, 1990), (b) motivationally consistent reconstruction and distortion of previous memories (e.g., McDonald & Hirt, 1997; Stillwell & Baumeister, 1997), and (c) increased implicit accessibility of motivationally consistent concepts in memory (Moskowitz, Gollwitzer, Wasel, & Schaal, 1999; Sinclair & Kunda, 1999). Further studies involving a variety of other types of motivations have demonstrated similar influences as well.

For example, in an illustration of selective recall and distortions of memory related to motivations for social connection, the more individuals typically adopt social goals that revolve around avoiding rejection, embarrassment, and bad feelings, the more negative and fewer positive social events

they recall and the more negatively they reconstruct those positive and neutral events they do recall (Strachman & Gable, 2006). Thus, individuals focused on avoiding rejection appear to selectively focus on this experience, presumably to vigilantly prepare themselves to prevent or cope with future failures of social connection. Similarly, the more individuals are typically motivated to remain autonomous in their relationships, the more they recall being less supportive than they actually were during a conflict with their relationship partner (Simpson, Rholes, & Winterheld, 2010). In contrast, the more individuals are typically motivated to continually pursue intimacy in their relationships, the more they recall feeling emotionally closer to their partner than they actually did during the conflict. Moreover, both of these conflict-related effects are particularly strong when individuals experience the conflict as distressing and their relationship concerns are more active. In another example of these type of effects that involve motivations for understanding and control, people selectively recall past events that imply they deserve a current outcome, thereby making the current outcome seem more under their control. That is, people who have just experienced a lucky break recall having performed more good deeds in the recent past, whereas those who have just experienced an unlucky break recall having performed more bad deeds in the recent past (Callan, Kay, Davidenko, & Ellard, 2009).

As with the effects of motivation on attention and information search described earlier, the effects of motivation on knowledge activation also operate at a more automatic and implicit level. When people's motivations for belonging are threatened after reliving an experience of social exclusion, they respond by spontaneously activating concepts related to the social groups to which they belong, as indicated by faster responses to these concepts in lexical decision tasks (Knowles & Gardner, 2008). Increased motivations for belonging can also increase the spontaneous activation of attitudes that match those of liked others. For example, participants interacting with an experimenter wearing a t-shirt that said "Eracism" showed weaker implicit activation of negative stereotypes associated with African Americans when this experimenter was friendly, but not when this experimenter was rude (Sinclair, Lowery, Hardin, & Colangelo, 2005). Motivations for control can have analogous effects as well. People whose feelings of control were challenged after viewing a video clip in which a person was stalked by a serial killer

spontaneously activated emotional concepts related to anger when viewing pictures of potentially hostile individuals and perceived greater anger in these individuals' objectively neutral expressions (Maner et al., 2005; see also DeWall & Baumeister, 2007).

EFFECTS ON ORGANIZATION OF CONCEPTS IN MEMORY

Finally, beyond affecting the activation of knowledge from memory, motivation for directional outcomes can influence how people come to organize this knowledge. Although less research has examined this phenomena, and only the effects of esteem-related motivations have been thoroughly investigated (but see Slotter & Gardner, 2009), several important outcomes have been identified. When motivations to sustain or protect esteem are activated, people (*a*) alter their self-concept to include attributes they believe will bring about successful outcomes (e.g., Klein & Kunda, 1992; Kunda & Santioso, 1989); (*b*) come to view the attributes they already possess as essential prerequisites for successful outcomes (Dunning, Leuenberger, & Sherman, 1995; Dunning, Perie, & Story, 1991; Kunda, 1987); and (*c*) redefine the criteria for successful outcomes or positive qualities in general so as to ensure that they meet these criteria (Beauregard & Dunning, 1998; Dunning & Cohen, 1992; see also Alicke, LoSchiavo, Zerbst, & Zhang, 1997).

In sum, motivations for directional outcomes can affect basic cognitive processes and influence thinking in several profound ways. These types of motivations influence not only how people search for, evaluate, and explain information but also how they activate, access, and organize their knowledge about themselves and others. The next section reviews research indicating that motivations for nondirectional outcomes can also have strong effects on information processing.

Influences of Motivations for Nondirectional Outcomes

Although there has been less research overall on motivation for nondirectional outcomes, some effects of this motivation have been considered in depth (e.g., Cacioppo, Petty, Feinstein, & Jarvis, 1996; Fiske & Neuberg, 1990; Kruglanski & Webster, 1996; Lerner & Tetlock, 1999). The two most prominent nondirectional motivations that have been studied are desires for *accuracy* (Fiske & Neuberg, 1990) and desires for simplicity and clarity, or *closure* (Kruglanski & Webster, 1996). As we

noted earlier, although accuracy and closure clearly represent particular outcomes that people are motivated to achieve during judgment, they are nondirectional outcomes because they do not dictate the specific content of the conclusions people want to reach. In the following sections, we consider the effects of these two motivations on many of the same cognitive processes as were discussed in the previous section.

Both accuracy and closure motivation have been studied in a variety of ways. Motivations for accuracy have been manipulated in terms of wanting to know as much as possible about a person on whom one is going to be dependent (Neuberg & Fiske, 1987), feelings of accountability for one's judgments (e.g., Tetlock, 1983), a "fear of invalidity" (e.g., Kruglanski & Freund, 1983), as well as simple desires to be as correct as possible (e.g., Neuberg, 1989). Motivations for closure have been manipulated in terms of time pressure (Kruglanski & Freund, 1983), desires to complete judgment tasks that are dull and unattractive (Webster, 1993), and desires to escape noisy environments (Kruglanski, Webster, & Klem, 1993; see Kruglanski & Webster, 1996). In the following discussion, we initially treat each of these varieties of accuracy or closure motivation as equivalent and then later consider some important differences among the effects of the various operationalizations.

EFFECTS ON ATTRIBUTION

In addition to specific biases stemming from esteem-, belonging-, and control-related motivations, research on attribution has identified more general biases, such as the tendency to fixate on one particular cause for some action or event while failing to adequately consider other possible alternatives (see Krull & Erickson, 1995). Although these attributional biases are frequently discussed in terms of the limits of cognitive capacity, there is evidence to suggest that they can be influenced by accuracy and closure motivations as well. For example, Tetlock (1985) demonstrated that, instead of the typical tendency to consider only one predominant explanation for others' observable behavior (i.e., that these individuals' general traits and attitudes were driving their actions), people with increased motivations for accuracy more evenly considered additional salient explanations (i.e., that temporary external circumstances could be influencing their actions). In contrast, Webster (1993) found that, when forming the same types of explanations

for behavior, increased motivations for closure had the opposite effect. People who wanted to quickly complete the attribution task and move on to a more desirable task displayed an even greater tendency to consider only one predominant explanation for others' behaviors than is typically found. Thus, motivations for accuracy versus closure can have opposite effects on people's attributional considerations of alternative causes (see Kruglanski & Webster, 1996).

EFFECTS ON INFORMATION SEARCH

Research on motivations for directional outcomes has demonstrated that people engage in prolonged information search when encountering evidence that disagrees with what they want to believe, and reduced information search when encountering evidence that agrees with what they want to believe. In contrast, accuracy motivation produces prolonged information search, and closure motivation produces reduced information search, regardless of whether the evidence encountered agrees or disagrees with prior personal beliefs.

For example, in one study on accuracy motivations, people instructed to form accurate impressions of an interview partner spent more time listening to their partner and provided more opportunities for this person to elaborate his or her opinions as compared to people given no special instructions (Neuberg, 1989). Moreover, these effects were found even when people began the interview with negative expectations (see also Maheswaran & Chaiken, 1991; for an extended review of related research see Chen & Chaiken, 1999; Eagly & Chaiken, 1993). More recent studies have further shown that people acting as the sole advisor for an upcoming decision to be made by someone else sought out more information about both the strengths and weakness of different choice alternatives (Jonas & Frey, 2003; see also Jonas, Schulz-Hardt, & Frey, 2005). These advisors also displayed equal attention toward either type of information, as compared to people deciding for themselves, who selectively viewed information about the strengths of their initially preferred choice. Most important, these differences between advisors and deciders were mediated by advisors' reports of greater concerns with accuracy in making the recommendations.

As an example of the effects of closure motivation on information search, in several studies people with varying motivations for closure worked with a partner to discuss the verdict of a mock trial based

on summarized legal analyses that were provided to them (Kruglanski et al., 1993). Prior to the discussion, individuals with stronger closure motivations expressed preferences for an easily persuadable partner, and, once the discussion began, they attempted to bring it to an end quickly and were more unwilling to consider alternative arguments. Similarly, in other studies in which people could perform unlimited trials of a perceptual task until they felt ready to make a judgment, individuals with stronger closure motivations requested fewer trials, even when the task was difficult and this decision jeopardized their performance (Roets, van Hiel, Cornelis, & Soetens, 2008). Indeed, recent studies have even demonstrated that individuals with stronger closure motivations experience greater feelings of threat and aversive physiological arousal when continuing to process information without reaching a definitive conclusion (Roets & van Hiel, 2008), and they experience greater regret when making choices that alter the status quo (thus removing closure; Mannetti, Pierro, & Kruglanski, 2007).

EFFECTS ON EVALUATION COMPLEXITY

In addition to affecting how long people evaluate and analyze information, motivations for nondirectional outcomes also influence the complexity of this analysis. Accuracy-motivated individuals form judgments that show greater consideration of ambiguity between conflicting opinions and evidence, whereas closure-motivated individuals show less consideration of ambiguity (Tetlock, 1983, 1998; see also Lerner & Tetlock, 1999). These findings have emerged from experiments in which people record their thoughts about topics such as affirmative action, American foreign policy, and the causes of certain historical events, which are then coded for the degree to which multiple perspectives on an issue are both identified and integrated into a framework that includes complex connections between them. Whether novices (e.g., college students) or experts (e.g., professional historians) on the issues they are analyzing, people with increased accuracy motivation provide more integrated and more complex analyses, whereas those with increased closure motivation provide less integrated and simpler analyses.

EFFECTS ON RECALL AND KNOWLEDGE ACTIVATION

Whereas motivations for directional outcomes have qualitative effects on what types of knowledge people recall and activate, motivations for

nondirectional outcomes have largely quantitative effects. Once again, accuracy motivation and closure motivation have opposite influences.

When motivated to form accurate impressions of others with whom they later expect to be partnered, people pay more attention to and remember more information about these individuals than when they do not expect any future interactions (Berscheid, Graziano, Monson, & Dermer, 1976; see also Srull, Lichtenstein, & Rothbart, 1985). However, individuals with dispositionally high need for closure spend less time reviewing information about others' behaviors and recall fewer of these behaviors (Dijksterhuis, van Knippenberg, Kruglanski, & Schaper, 1996). In addition, when forming impressions of others, people with increased motivations for accuracy activate more unique, or *individuating*, trait and behavioral information (Kruglanski & Freund, 1983; Neuberg & Fiske, 1987), whereas people with increased motivations for closure rely instead on common information associated with stereotypes about members of particular social categories (Dijksterhuis et al., 1996; Kruglanski & Freund, 1983; see also Moskowitz, 1993). These individuation effects can even operate at a perceptual level; although people typically show a *cross-race effect*, in which they display better recognition for faces of their own versus another race, individuals with increased motivations to accurately individuate other-race faces do not differ in their recognition of same-race versus other-race faces (Hugenberg, Miller, & Claypool, 2007).

Beyond influencing the recall and activation of knowledge from memory, accuracy and closure motivations also affect people's application of this activated knowledge during judgment. Typically, concepts or attitudes that are recently or frequently activated create biases such that people either directly assimilate to or contrast from accessible constructs without considering alternative perspectives (see Higgins, 1996). Increased accuracy motivations attenuate these biases by increasing the activation of alternative interpretations. Thus, when evaluating behaviors that are somewhat ambiguous, people with strong accuracy motivations are less likely to interpret these behaviors solely based on whatever personality traits are most accessible (Thompson et al., 1994; see also Dijksterhuis, Spears, & Lepinasse, 2001; Schuette & Fazio, 1995).

In contrast to this effect of accuracy motivations, increased closure motivations exacerbate accessibility-related biases by decreasing the activation of alternative interpretations (see Kossowska, 2007). People

are more likely to interpret ambiguous behaviors based on accessible personality traits when their closure motivation is high (Ford & Kruglanski, 1995; see also Sanbonmatsu & Fazio, 1990). In addition, people with stronger motivations for closure are more likely to falsely "transfer" their accessible memories of and feelings toward individuals they know well to a new individual who shares a few of the same qualities as the known individuals (Pierro & Kruglanski, 2008; see also Lun, Sinclair, Whitchurch, & Glenn, 2007). Finally, stronger motivations for closure can even influence the extent to which people display broad cultural differences in judgment and behavior. Several different programs of research have suggested that the general mindsets for judgment and information processing that people develop from living in different cultures can function like accessible knowledge structures (see Hong, Morris, Chiu, & Benet-Martínez, 2000); closure motivations thus again increase people's application of this accessible knowledge and their conformity to the cultural norms that are currently most predominant (Chiu, Morris, Hong, & Menon, 2000; Fu et al., 2007; Kosic, Kruglanski, Pierro, & Mannetti, 2004; see Kruglanski, Pierro, Mannetti, & De Grada, 2006).

Overall, then, motivations for nondirectional outcomes can also affect basic cognitive processes and profoundly influence thinking. Whereas motivations for directional outcomes were earlier shown to alter *how* people activate, evaluate, and explain information during reasoning, motivations for non-directional outcomes (at least in terms of the accuracy and closure motivations reviewed here) instead alter *how much* activation, evaluation, or explanation, in fact, occurs. Furthermore, as the findings just reviewed illustrate, such quantitative differences in thought can often affect the overall outcomes of people's judgments and decisions just as much as the qualitative differences described earlier.²

Limits to Outcome-Motivated Thinking

Although people have an impressive array of cognitive mechanisms at their disposal when pursuing desired outcomes during judgment, limits do exist concerning when they engage these mechanisms. These limits are first described for directional thinking and then for nondirectional thinking.

REALITY CONSTRAINTS ON MOTIVATIONS FOR DIRECTIONAL OUTCOMES

Despite any specific outcomes for which people have some preference for during judgment, most

individuals still acknowledge the existence of some kind of “objective reality” about whatever information they are considering. That is, motivated thinking related to directional outcomes operates within what Kunda (1990) termed *reality constraints*. Therefore, although people may choose motivationally consistent attributions, engage in selective recall, or criticize motivationally inconsistent evidence, they are not entirely unresponsive to world around them (except perhaps in extreme circumstances; see Bachman & Cannon, Chapter 34).

Indeed, the meta-analysis by Hart et al. (2009) that showed a sizable effect for selective attention toward motivationally consistent information also showed a significant effect for information quality as well. That is, compared to information that was diagnostic and relevant for people’s judgments, non-diagnostic and irrelevant information did not receive as much attention, even if this latter information was motivationally consistent. Moreover, if the only motivationally consistent information available was of low quality, people actually showed greater attention toward motivationally inconsistent information. Thus, although a simultaneous comparison of the independent effects of motivation versus information quality on selective attention revealed that the motivational effects were larger, information quality still had a significant influence beyond people’s motivations, as would be expected if people are generally sensitive to reality constraints.

Some studies have even gone a step further and begun to outline circumstances in which reality constraints are likely to have more or less influence. One clear finding from these studies is that people’s judgments are guided by motivations for directional outcomes to a greater extent when what they are interpreting or judging involves some vagueness or ambiguity (e.g., Dunning, Meyerowitz, & Holtzberg, 1989; Hsee, 1995). For example, when there is more potential for constructing idiosyncratic criteria for a particular attribution (e.g., judging whether one possesses somewhat vague traits like insecurity), then people select criteria that best allow them to reach their desired conclusion. However, when there is less potential for this construction (e.g., judging whether one possesses more precise traits such as punctuality), people engage in less motivated reasoning (Dunning et al., 1989). Also, when there is ambiguity concerning others’ expressions and identities or one is forming impressions of strangers with whom one has not interacted, there is greater potential for motivationally consistent

projection of emotion or potential friendliness (Maner et al., 2005). However, when such ambiguity does not exist, the likelihood of motivationally consistent projection decreases. For example, when motivations for belonging are threatened by experiences of rejection, people do not attribute greater friendliness to the individuals responsible for this exclusion, whose feelings toward them are anything but ambiguous (Maner et al., 2007). Overall, these results suggest that thinking and reasoning inspired by directional outcomes does not so much lead people to ignore the sometimes disappointing reality they face as it inspires them to exploit in their favor whatever uncertainties there are in this reality.

COGNITIVE-RESOURCE CONSTRAINTS ON ACCURACY MOTIVATION

Most of the effects of accuracy motivation reviewed earlier were driven by increased information processing during judgment. Therefore, the effects of strong accuracy motivation should be reduced under circumstances in which people’s ability to engage in such information processing is constrained (Fiske & Neuberg, 1990). Indeed, Pendry and Macrae (1994) showed that whereas accuracy-motivated individuals increased their use of information about individuating traits in judging others’ behaviors when they possessed their full information-processing resources, as described earlier (see Neuberg & Fiske, 1987), accuracy-motivated individuals whose processing resources were depleted based their impression primarily on categorical information, similar to those who had little accuracy motivation (see also Kruglanski & Freund, 1983). In addition, Sanbonmatsu and Fazio (1990) demonstrated that the reduction in people’s assimilation of their judgments to highly accessible attitudes associated with accuracy motivation disappears when people are placed under time pressure that prevents extended information processing (see also Roets et al., 2008).

THE TENUOUS LINK BETWEEN ACCURACY MOTIVATION AND ACCURATE REASONING

Another important consideration regarding the effects of accuracy motivation on thinking and reasoning is that even when people high in accuracy motivation can engage in extended information processing, this does not guarantee that they will make more accurate judgments. At times, evidence beyond what is immediately and effortlessly available does not exist or has faded from memory (see,

e.g., Thompson et al., 1994). Moreover, people can be influenced by biases of which they are either unaware or uncertain how to correct. In these circumstances, although accuracy motivation might increase information search, recall, and consideration of multiple interpretations, it would not be expected to eliminate judgment errors (Fischhoff, 1982), and it could even increase them (Pelham & Neter, 1995; Tetlock & Boettger, 1989).

DISTINCTIONS AMONG CIRCUMSTANCES THAT LEAD TO ACCURACY MOTIVATION

As alluded to earlier, the different types of accuracy-motivation inductions reviewed here are not always equivalent and can have markedly different effects. For example, although having one's outcomes dependent on another person can increase desires for accuracy in diagnosing that person's true character (e.g., Neuberg & Fiske, 1987), in other cases, such circumstances can produce a desire to see a person upon whom one is depending in the best possible light (e.g., Berscheid et al., 1976; Klein & Kunda, 1992; see Kruglanski, 1996). Also, although believing that one's judgment has important consequences may motivate an accurate consideration of all the relevant evidence, it could also motivate more general increases in information processing that is not necessarily focused on accuracy (see note 2; Petty & Wegener, 1999). Finally, although justifying one's judgments to others can motivate accuracy when the preferences of the audience are unknown, it can also produce motivations for more directional outcomes, such as ingratiation toward this audience, when the opinion of the audience is known (Jonas et al., 2005; Lerner & Tetlock, 1999; Tetlock, 1983). Therefore, when attempting to anticipate the effects of accuracy motivation on reasoning in a particular situation, it is important to consider both the current source of this motivation and the larger context in which it occurs.

THE INFLUENCE OF INFORMATION AVAILABILITY ON CLOSURE MOTIVATION

Certain qualifications should also be noted regarding the effects of closure motivation. The findings reviewed earlier involved the tendency for people with strong closure motivation to quickly assimilate their judgments to readily available or highly accessible information, leading to an early "freezing" of their information search. However, in situations in which there is little information available, strong closure motivation may inspire efforts

to find something clear and concise to "seize" upon and *increase* information search (see Kruglanski & Webster, 1996). For example, individuals with stronger closure motivation prefer easily persuadable partners and are unwilling to consider alternative arguments only when they have enough information at their disposal to form some clear initial impression; without such information these individual express a desire for partners who are highly *persuasive* and quickly shift toward their partner's point of view (Kruglanski et al., 1993). Similarly, when individuals with strong closure motivation are forced to adopt roles that are culturally unfamiliar, they again display an increase rather than a decrease in information search (Fu et al., 2007).

Conclusions on Outcome-Motivated Thinking

Recent research has uncovered many potential routes by which people's desires for particular judgment outcomes can affect their thinking and reasoning. To summarize, a number of basic cognitive processes during reasoning are influenced by both motivations for directional outcomes, where people have a specific, content-dependent conclusion they want to reach, and motivations for nondirectional outcomes, where people's preferred conclusions are more general and content independent. These include (a) the explanation of events and behaviors; (b) the organization, recall, and activation of knowledge in memory; and (c) the pursuit and evaluation of evidence relevant to decision making. Outcome-motivation effects involve both how such cognitive processes are initiated and directed, as well as how thoroughly these processes are implemented. Moreover, in any given situation, the specific cognitive processes influenced by outcome motivation are typically those that aid the gathering and interpretation of information supporting the favored outcome. In this self-fulfilling way, people's outcome-motivated reasoning often brings about their desired conclusions.

Strategy-Motivated Thinking

Although outcome-motivated thinking has been the most widely studied form of motivated reasoning, there are other varieties of motivational influences on cognition. One alternative perspective suggests that people are motivated not only with respect to the outcomes of their judgments but also with respect to the manner in which they make these judgments. That is, not only do people have

preferred conclusions, but they also have *preferred strategies* for reaching their conclusions (Higgins and Molden, 2003; cf. Tyler & Blader, 2000). Therefore, independent of whatever outcome holds the most interest for them, people may be motivated to reach these outcomes using strategies that “feel right” and sustain their current motivational orientation. Several lines of research have recently examined how motivations for particular judgment strategies can also influence people’s basic cognitive processes. In the vast majority of these studies, strategic motivations were measured and manipulated in one of two ways: in terms of people’s *regulatory focus* (see Higgins, 1997) or their *regulatory mode* (Kruglanski et al., 2000).

Regulatory focus describes the distinction between two basic motivational orientations: those involving *promotion concerns* focused on advancement and approaching gains versus avoiding nongains, and those involving *prevention concerns* focused on security and approaching nonlosses versus avoiding losses. Because they revolve around advancement, promotion concerns create preferences for *eager strategies* emphasizing a broad, open-minded search for opportunities for gain, even at the risk of committing errors and accepting losses. In contrast, because they revolve around security, prevention concerns create preference for *vigilant strategies* emphasizing a more narrow, close-minded focus on protecting against loss, even at the risk of missing opportunities for potential gain (Higgins, 1997; Molden, Lee, & Higgins, 2008).

Regulatory mode describes the distinction between two additional motivational orientations: those involving *locomotion* concerns with initiating action and continual, smooth movement toward achieving goals by whatever means available, versus those involving *assessment* concerns with ensuring a critical analysis of the most effective and appropriate means of goal pursuit before any action is taken. Because they revolve around movement, locomotion concerns create preferences for strategies of successive elimination and moving from state to state to sustain a feeling of continuous movement. In contrast, because they revolve around analysis, assessment concerns create preferences for strategies of exhaustive comparison and delayed responses to sustain a feeling of thoroughness and a focus what is optimal (Higgins, Kruglanski, & Pierro, 2003; Kruglanski et al., 2000).

Thus, even in circumstances in which individuals are pursuing the same outcome (e.g., achieving the

best performance possible), they may show significant differences in their pursuit of this outcome depending upon whether they are currently promotion- versus prevention-focused (e.g., by trying any available means that might possibly lead to success versus using only the few means that would most probably lead to success, respectively), or whether they are locomotion- versus assessment-focused (e.g., by concentrating on continually taking whatever actions they can to maintain movement toward achievement versus thoroughly and critically evaluating what specific actions are most likely to boost achievement and should be the main focus their efforts, respectively). The studies reviewed next illustrate the general effects of these types of strategic motivations on several aspects of thinking and reasoning that are similar to those reviewed in the previous sections (for larger overviews, see Kruglanski, Orehhek, Higgins, Pierro, & Shalev, 2010; Molden et al., 2008).

Influences of Motivated Strategies of Judgment

EFFECTS ON THE CONSIDERATION OF ALTERNATIVE HYPOTHESES

Considering alternative hypotheses is a fundamental component of many varieties of thinking. As alluded to earlier, people’s strategic motivations can influence their preferred way of evaluating these alternatives. First, the eager, open-minded strategies motivated by promotion concerns should lead people to broadly consider multiple alternatives so as to increase the chance of finding the correct answer. In contrast, the vigilant, close-minded strategies motivated by prevention concerns should lead people to narrow their consideration to fewer alternatives so as to increase the chance of rejecting incorrect answers (see Molden et al., 2008). Thus, when forming impressions, promotion-focused individuals should tend to consider and endorse a greater number of possibilities than prevention-focused individuals.

These effects were illustrated in several studies by Liberman, Molden, Idson, and Higgins (2001). One important instance of considering alternatives occurs when people form hypotheses about what they are perceiving (see Hegarty & Stull, Chapter 31); when asked to identify vague and distorted objects in a series of photographs, promotion-focused individuals generated a greater number of alternatives than prevention-focused individuals (see also Friedman & Förster, 2001). Another important instance of considering alternatives, discussed previously, occurs when people form attributions about

their own and others' actions (Kelley, 1973); when asked to evaluate explanations for a target person's behavior, promotion-focused individuals again endorsed a greater number of alternatives than prevention-focused individuals. Molden and Higgins (2008) found similar effects when people explained their own behavior following success or failure at an intellectual task. Yet another important instance of considering alternatives occurs when people assign individuals and objects to general categories (see Rips, Medin, & Smith, Chapter 11); Molden and Higgins (2004) found that when asked to categorize the actions of particular individuals, promotion-focused individuals once again generated more alternatives than prevention-focused individuals.

Several of these studies by Molden and colleagues further showed that motivational differences in people's consideration of alternatives had significant consequences for the final impressions they formed (Liberman et al., 2001; Molden & Higgins, 2008). In general, the more alternative explanations people consider for a particular outcome, the less confident they can be in the validity of any single explanation (see Kelley, 1973). Thus, promotion-focused individuals ultimately formed more equivocal impressions and displayed less generalization of their attributions to future situations than prevention-focused individuals.³

There is also evidence that locomotion versus assessment concerns can influence people's consideration of alternatives (Avnet & Higgins, 2003). As noted earlier, locomotion concerns create preferences for strategies of progressive elimination, whereas assessment concerns create preferences for strategies of exhaustive comparison. Supporting these preferences, individuals whose locomotion concerns were activated later showed greater satisfaction with a decision when they used a progressive elimination strategy (i.e., successively making comparisons among choice alternatives based on one attribute dimension at a time and on each round eliminating the worst option for that attribute) than when they used an exhaustive comparison strategy (i.e., simultaneously comparing all of the alternatives on all of the attribute dimensions). Individuals whose assessment concerns were activated showed an opposite pattern of satisfaction.

Overall, then, strategic preferences play a substantial role in people's generation of and selection among alternatives during a number of important cognitive processes. Furthermore, several of the studies reviewed earlier also included measures of

people's motivations for more general outcomes such as accuracy and closure, and the effects of regulatory focus and regulatory mode remained when statistically controlling for these outcome motivations. Moreover, whereas the effects of accuracy or closure motivations on people's consideration of alternatives are typically driven by differences in the time or effort people invest in gathering or considering these alternatives during decision making, the effects of concerns with promotion versus prevention or locomotion versus assessment were instead typically driven more by differences in how people weighted and selected among those alternatives that they have gathered (see e.g., Liberman et al., 2001). Such findings thus support the classification of strategic preferences as a separate source of motivated thinking.

EFFECTS ON COUNTERFACTUAL THINKING

Besides generating and evaluating hypotheses, another way in which people consider alternatives during reasoning is in their use of *counterfactuals*. Counterfactual thinking involves mentally undoing the present circumstances by pondering what would have happened if only different decisions had been made or different actions taken (Roese, 1997). Many studies have characterized different varieties of counterfactual thinking, but one broad distinction involves *additive* counterfactuals concerning the reversal of previous inaction (if only I had acted, things might have gone better), versus *subtractive* counterfactuals concerning the reversal of previous actions (if only I had not acted, things wouldn't be so bad).

Because additive counterfactuals allow the mental correction of missed opportunities for gain, this type of thinking represents a more eager strategy of considering alternative realities. In contrast, because subtractive counterfactuals allow the correction of mistakes that resulted in loss, this type of thinking represents a more vigilant strategy of considering alternate realities. Consistent with this logic, both when analyzing hypothetical examples and when describing particular instances of their own behavior, promotion-focused individuals offer more additive counterfactuals, whereas prevention-focused individuals offer more subtractive counterfactuals (Roese, Hur, & Pennington, 1999; see also Molden, Lucas, Gardner, Dean, & Knowles, 2009). Research on counterfactual thinking has traditionally assumed that subtractive counterfactuals are more common than additive counterfactuals and that failures associated with action inspire more regret than failures

associated with inaction (Roese, 1997). However, the results of these studies demonstrate that people's strategic preferences can alter these typical circumstances, leading additive counterfactuals to be more common for promotion-focused individuals and to cause these individuals greater regret (see also Camacho, Higgins, & Lugar, 2003).

Whereas strategic preferences associated with promotion or prevention concerns influence the type of counterfactuals people generate, strategic preferences associated with locomotion or assessment concerns influence how many counterfactuals people generate overall. In general, counterfactual thinking functions as a means of analyzing past decisions to ensure better decisions in the future (Roese, 1997). Therefore, increased motivations to critically evaluate one's decision-making processes should increase counterfactual thinking, whereas increased motivations to "move on" after making a decision should decrease counterfactual thinking. Consistent with this logic, both when considering hypothetical examples and when describing particular instances of their own behavior, individuals with stronger assessment concerns generated more counterfactuals and displayed stronger experiences of regret, whereas individuals with stronger locomotion concerns generated fewer counterfactuals and displayed less regret (Pierro et al., 2008).

EFFECTS ON TRADE-OFFS BETWEEN FAST AND ACCURATE INFORMATION PROCESSING

A major question across many areas of psychology has been when and why people emphasize either speed or accuracy during decision making (e.g., Josephs & Hahn, 1995; Zelaznik, Mone, McCabe, & Thaman, 1988). Förster, Higgins, and Bianco (2003) investigated whether promotion-focused preferences for eager strategies would result in a priority for faster information processing and a higher quantity of output in a search for possible opportunities for gain, whereas prevention preferences for strategic vigilance would result in a priority for more accurate information processing and a higher *quality* of output in an effort to protect against loss. That is, analogous to the findings of Molden and colleagues described earlier (Liberman et al., 2001; Molden & Higgins, 2004, 2008), promotion-focused individuals might selectively pursue faster information processing to increase the number of correct answers they might identify, whereas prevention-focused individuals might selectively pursue more accurate information processing to

increase the number of incorrect answers they might eliminate. On several different motor-performance and information-processing tasks, individuals with either a chronic or temporarily activated focus on promotion did indeed perform these tasks faster, but with lower accuracy, whereas individuals with a chronic or temporarily activated focus on prevention performed more slowly but with greater accuracy. Moreover, these effects of strategic preferences were stronger at the end than at the beginning of the task, showing the classic "goal looms larger" effect in which motivational intensity increases as people move closer to goal completion (see Lewin, 1935).

Concerns with locomotion versus assessment can also affect the priority given to fast or accurate information processing. Mauro, Pierro, Mannetti, Higgins, and Kruglanski (2009) showed that when performing a group decision-making task, the preferences for progress and movement held by those focused on locomotion resulted in faster, but less accurate decisions, whereas the preferences for prolonged, careful analysis held by those focused on assessment resulted in slower but more accurate decisions (see also Mannetti et al., 2009). Interestingly, results further showed that when both preferences were represented in groups featuring equal numbers of locomotion- and assessment-focused individuals, decisions were made with a high level of accuracy without sacrificing speed.

EFFECTS ON KNOWLEDGE ACTIVATION AND RECALL

Analogous to the selective recall and activation of information from memory inspired by motivations for directional outcomes, another influence of strategic preferences is to increase sensitivities to, and recall of, information that is particularly relevant to these preferences. In one study, after reading an essay about the life of a hypothetical target person in which this person used both eager, advancement-oriented strategies (e.g., waking up early in order to be on time for a favorite class) and vigilant, protection-oriented strategies (e.g., being careful not to sign up for a class schedule that conflicted with other desired activities), promotion-focused individuals showed greater recall for events involving the use of eager as compared to vigilant strategies, whereas prevention-focused individuals showed the reverse effect (Higgins, Roney, Crowe, & Hymes, 1994). Similarly, in another study, promotion-focused individuals showed greater recall for a target person's experiences of both the presence and absence of gains

(e.g., finding \$20 on the street or missing a movie that he wanted to see, respectively), whereas prevention-focused individuals showed greater recall for a target person's experiences of both the presence and absence of losses (e.g., being stuck in a crowded subway for an extended period of time or getting a day off from a particularly arduous class schedule, respectively; Higgins & Tykocinski, 1992). Future research should examine whether locomotion or assessment concerns can have the same types of effects.

STRATEGIC PREFERENCES AND REGULATORY FIT

The studies reviewed thus far have demonstrated that people's motivational concerns lead them to prefer and choose particular types of judgment strategies. However, the extent to which people can follow their preferences can vary by their circumstances. For example, some situations may generally require greater use of one type of strategy versus another, such as when one supervisor demands constant innovation and the pursuit of multiple approaches for a particular goal, whereas another supervisor demands the use of proven techniques and careful analysis of how to pursue the goal. Given such situational variability, a major focus of research on strategic preferences has been on whether the strategy demanded by a situation fits or does not fit individuals' current motivational concerns (see Higgins, 2000, 2008).

Although space limitations prevent a thorough review of the research on such *regulatory fit* here, the general findings have been that regulatory fit increases the perceived value of the goal one is pursuing and strengthens engagement during goal pursuit. That is, whether arising from strategies that fit promotion versus prevention concerns or locomotion versus assessment concerns, circumstances that create regulatory fit (as compared to nonfit) lead people to "feel right" about their goal pursuit, engage more strongly, and react more intensely to goal success or failure. When regulatory fit exists, people thus (*a*) see the strategies they are using as the right way to approach goal pursuit (Camacho et al., 2003; Pierro, Presaghi, Higgins, & Kruglanski, 2009), (*b*) show greater enjoyment of goal-directed action (Freitas & Higgins, 2002; Pierro, Kruglanski, & Higgins, 2006); and (*c*) assign greater worth to the outcomes they are pursuing (Avnet & Higgins, 2003; Higgins, Idson, Freitas, Spiegel, & Molden, 2003). More recently, research has also explored the broader consequences of regulatory fit experiences for judgment and

information processing. This research has further shown that, as compared to experiences of nonfit, experiences of fit (*a*) facilitate a more flexible, active, and exploratory response strategy (due to the increased enjoyment and valuing of goal-directed action, see Maddox & Markman, 2010) and (*b*) encourage reduced information search and an increased application of whatever knowledge happens to be particularly accessible in memory (because this knowledge "feels right," see Koenig, Cesario, Molden, Kosloff, & Higgins, 2010; Vaughn et al., 2006).

Conclusions on Strategy-Motivated Thinking

In sum, several continuing programs of research have now demonstrated that, beyond the effects on reasoning of people's desires for particular judgment outcomes, there are additional effects on reasoning of people's desires to use particular judgment strategies. Whether strategic preferences arise from concerns with promotion versus prevention or locomotion versus assessment, these preferences influence (*a*) the generation and testing of hypotheses, (*b*) the occurrence of counterfactual thinking, (*c*) the prioritization of fast versus accurate information processing, (*d*) knowledge activation and recall, and (*e*) cognitive flexibility during judgment and decision making. Like outcome-motivation effects, strategy-motivation effects include what cognitive processes are initiated, how thoroughly these processes are implemented, and how much the resulting information that is gathered is valued in a final decision.

Conclusions and Future Directions

The sheer number and diversity of the studies reviewed here is a testament to the growing recognition of the importance of motivational perspectives on cognition. The richness and consistency of the findings emerging from these studies is also a testament to the utility of this perspective in the study of thinking and reasoning. We optimistically forecast a continued expansion of research informed by motivational perspectives and, in conclusion, briefly outline several directions we believe should be priorities for the future.

The first direction involves a greater exploration of how, beyond the common effects of the various outcome- or strategy-related motivations reviewed here, different types of directional or strategic motivations may have unique effects. For example, although we have discussed how motives for esteem, belonging, and control all generally influence the selective recall

of motivationally relevant information, given that these motives all represent fundamentally different needs, there might be additional processes that each motive affects in different ways (cf., Maner, 2009). Some preliminary evidence for these types of distinct effects already exists. People respond to interaction partners differently when motives for belonging and control have both been threatened rather than either one of these on its own (Warburton, Williams, & Cairns, 2006). People also use fundamentally different strategies in their attempts to compensate for esteem threats versus belonging threats (Knowles, Lucas, Molden, Gardner, & Dean, 2010). Future studies should thus continue to better define both the commonalities and differences in the effects of various motivations on cognition.

A second direction for future research involves further investigation of the interaction between different types of motivations in thought and behavior. For example, De Dreu, Nijstad, and van Knippenberg (2008) have begun to analyze group decision making in terms of the joint effects of directional motivations involving people's desires for more self-focused or other-focused outcomes and nondirectional motivations involving their desires for closure. They argue that when closure motivation is high, people pursue low effort means of attaining self-focused or other-focused outcomes (e.g., laziness and inaction versus conformity to a group position, respectively), but that when closure motivation is low (or accuracy motivation is high) people use more elaborate means to attain these outcomes (e.g., advocating one's own position and disagreeing with others versus attempting to reconcile differing opinions, respectively). Because at any given time people are rarely focused on only achieving a single goal, this focus on the implications of multiple active motivations is an important expansion of existing perspectives and is perhaps a better reflection of how people experience the influence of different motivations in the judgments they make in their daily lives (Fishbach & Ferguson, 2007). Additional research could also explore how, beyond particular types of motivations that interact, there could also be other types of motivations that are incompatible and inhibit each other when active (e.g., motivations for dominance and control versus motivations for egalitarianism and equality or motivations for order versus motivations for freedom and independence; see Maio, Pakizeh, Cheung, & Rees, 2009).

One final direction for future research is the synthesis of research on how motivation influences

reasoning with research on how affect influences reasoning (e.g., Forgas, 2001; Martin & Clore, 2001). Many of the changes in the quality and quantity of information processing found in research on emotion and cognition bear a striking resemblance to the motivational effects reviewed here. For example, positive moods have generally been found to support less thorough and complex information processing, similar to closure motivation, whereas negative moods have generally been found to support more thorough and complex information processing, similar to accuracy motivation (for a review, see Schwarz & Clore, 2007). In addition, a recent meta-analysis on the range of cognitive effects of happy or anxious emotions has related these effects to the strategic motivations inspired by promotion or prevention concerns (Baas, De Dreu, & Nijstad, 2008). Indeed, Lucas, Molden, and Gardner (unpublished data) have even recently found evidence that the common effects of a wide variety of anxiety-provoking experiences (e.g., fears of contamination, worries about social rejection, discomfort with uncertainty) on more vigilant and avoidant mindsets is partially mediated by the fact that these experiences all evoke prevention-focused strategic motivations (see also Higgins, 1997). We should emphasize that we are not proposing that the instances of motivated thinking reviewed here are ultimately just rooted in emotion, as many of the studies reviewed carefully controlled for affective influences and continued to find independent effects. Rather, we are suggesting that it would be fruitful to further investigate both how affective thinking may give rise to motivational thinking and how motivational thinking may give rise to affective thinking (see also Higgins, 2001).

To conclude, this chapter has reviewed research that displays the broad applicability of motivational perspectives to the study of thinking and reasoning. Through this review, we have attempted to convey the potential utility of these perspectives and to advocate a greater incorporation of principles of outcome- and strategy-based motivation in future research. The further refinement and elaboration of these principles, we believe, will benefit not only the study of judgment and reasoning but also cognitive science in general.

Notes

1 One literature notably absent from our review concerns affective and emotional influences on reasoning. This important, and extensive, literature certainly enjoys a central place in the study of motivated thinking. However, the topic of affect and cognition has been the subject of several entire handbooks on

its own (seeForgas, 2001; Martin & Clore, 2001), not to mention its own separate journal (*Cognition and Emotion*). Therefore, rather than attempt a limited overview of this major topic alongside the others mentioned earlier, we instead refer the interested reader to these other sources. The larger relation between research on emotional thinking and the research described here is discussed briefly later in the chapter.

2 Another type of nondirectional outcome motivation that has been the focus of considerable study is the *need for cognition*, or a general desire for elaborative thinking and increased cognitive activity (Cacioppo et al., 1996). At times, the need for cognition has been considered equivalent to accuracy motivation (Chen & Chaiken, 1999). Consistent with this perspective, research has shown that an increased need for cognition can affect thinking in the same way as heightened accuracy motivation, reducing biases during attribution (D'Agostino & Fincher-Kiefer, 1992), increasing recall (Srull et al., 1985), lessening assimilation to highly accessible attitudes (Florack, Scarabis, & Bless, 2001), and increasing information search (Verplanken, 1993). However, at times the effects of the need for cognition differ from those of accuracy motivation. Accuracy motivation, because it inspires a thorough consideration of *all* available evidence, weakens the tendency to base judgments on early superficial impressions (i.e., primacy effects; Kruglanski, & Freund, 1983). In contrast, the need for cognition, because it simply inspires cognitive elaboration even if this involves only part of the available evidence, can lead to increased rumination on one's early superficial impressions and *strengthen* primacy effects (see Petty & Wegener, 1999). Given these conceptual and empirical distinctions, we have not included research on need for cognition in our larger review of the effects of accuracy motivation and consider it a separate form of nondirectional outcome motivation (for a review of need for cognition effects, see Cacioppo et al., 1996).

3 Although the studies reviewed consistently show that promotion-focused individuals typically consider more alternatives than prevention-focused individuals during judgment, several other studies have identified circumstances that produce an important reversal of this effect (Scholer, Stroessner, & Higgins, 2008). When specifically contemplating possible threats or identifying negative outcomes, vigilant strategies of protecting against loss actually require a broader consideration of any alternatives that might signal the presence of a threat, whereas eager strategies of seeking gains are less focused on the potential for threat. Therefore, in these circumstances prevention-focused individuals consider more alternative threats and negatively valenced pieces of information than promotion-focused individuals (see also Scholer, Zou, Fujita, Stroessner, & Higgins, 2010; cf. Mishra & Lalumière, 2010).

References

- Alicke, M. D., LoSchiavo, F. M., Zerbst, J. I., & Zhang, S. (1997). The person who outperforms me is a genius: Maintaining perceived competence in upward social comparison. *Journal of Personality and Social Psychology*, 73, 781–789.
- Allport, F. H. (1920). The influence of the group upon association and thought. *Journal of Experimental Psychology*, 3, 159–182.
- Avnet, T., & Higgins, E. T. (2003). Locomotion, assessment, and regulatory fit: Value transfer from "how" to "what". *Journal of Experimental Social Psychology*, 39, 525–530.
- Baas, M., De Dreu, C. K. W., & Nijstad, B. A. (2008). The mood-creativity link reconsidered: A meta-analysis of 25 years of research. *Psychological Bulletin*, 134, 779–806.
- Balceris, E., & Dunning, D. (2007). Cognitive dissonance and the perception of natural environments. *Psychological Science*, 18, 917–921.
- Balceris, E., & Dunning, D. (2010). Wishful seeing: Desired objects are seen as closer. *Psychological Science*, 21, 147–152.
- Baumeister, R. F. (1998). The self. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 1, pp. 680–740). New York: McGraw-Hill.
- Baumeister, R. F., & Leary, M.R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117, 497–529.
- Beauregard, K. S., & Dunning, D. (1998). Turning up the contrast: Self-enhancement motives prompt egocentric contrast effects in social judgment. *Journal of Personality and Social Psychology*, 74, 606–621.
- Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, 74, 183–200.
- Berscheid, E., Graziano, W., Monson, T., & Dermer, M. (1976). Outcome dependency: Attention, attribution, and attraction. *Journal of Personality and Social Psychology*, 34, 978–989.
- Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, 119, 197–253.
- Callan, M. J., Kay, A. C., Davidenko, N., & Ellard, J. H. (2009). The effects of justice motivation on memory for self- and other-relevant events. *Journal of Experimental Social Psychology*, 45, 614–623.
- Camacho, C. J., Higgins, E. T., & Lugar, L. (2003). Moral value transfer from regulatory fit: What feels right *is* right and what feels wrong *is* wrong. *Journal of Personality and Social Psychology*, 84, 498–510.
- Chen S., & Chaiken, S. (1999). The heuristic-systematic model in its broader context. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 73–96). New York: Guilford Press.
- Chiu, C. Y., M. W. Morris, Y. Y. Hong, & Menon, T. (2000). Motivated cultural cognition: The impact of implicit cultural theories on dispositional attribution varies as a function of need for closure. *Journal of Personality and Social Psychology*, 78, 247–259.
- D'Agostino, P. R., & Fincher-Kiefer, R. (1992). Need for cognition and the correspondence bias. *Social Cognition*, 10, 151–163.
- De Dreu, C. K. W., Nijstad, B. A., & Van Knippenberg, D. (2008). Motivated information processing in group judgment and decision making. *Personality and Social Psychology Review*, 12, 22–49.
- DeWall, C. N., & Baumeister, R. F. (2007). From terror to joy: Automatic tuning to positive affective information following mortality salience. *Psychological Science*, 18, 984–990.
- DeWall, C. N., Maner, J. K., & Rouby, D. A. (2009). Social exclusion and early-stage interpersonal perception: Selective attention to signs of acceptance. *Journal of Personality and Social Psychology*, 96, 729–741.
- Dijksterhuis, A., van Knippenberg, A., Kruglanski, A. W., & Schaper, C. (1996). Motivated social cognition: Need for closure effects on memory and judgment. *Journal of Experimental Social Psychology*, 32, 254–270.
- Dijksterhuis A., Spears R., & Lepinasse V. (2001). Reflecting and deflecting stereotypes: Assimilation and contrast in

- impression formation and automatic behavior. *Journal of Experimental Social Psychology*, 37, 286–299.
- Ditto, P. H., Scipensky, J. A., Munro, G. D., Apanovich, A. M., & Lockhart, L. K. (1998). Motivated sensitivity to preference-inconsistent information. *Journal of Personality and Social Psychology*, 75, 53–69.
- Dunning, D. (1999). A newer look: Motivated social cognition and the schematic representation of social concepts. *Psychological Inquiry*, 10, 1–11.
- Dunning, D., & Cohen, G. L. (1992). Egocentric definitions of traits and abilities in social judgment. *Journal of Personality and Social Psychology*, 63, 341–355.
- Dunning, D., Leuenberger, A., & Sherman, D. A. (1995). A new look at motivated inference: Are self serving theories of success a product of motivational forces? *Journal of Personality and Social Psychology*, 69, 58–68.
- Dunning, D., Meyerowitz, J. A., & Holtzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology*, 57, 1082–1090.
- Dunning, D. A., Perie, M., & Story, A. L. (1991). Self-serving prototypes of social categories. *Journal of Personality and Social Psychology*, 61, 957–968.
- Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. New York: Harcourt Brace College Publications.
- Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty* (pp. 237–262). New York: Cambridge University Press.
- Fishbach, A., & Ferguson, M. F. (2007). The goal construct in social psychology. In A. W. Kruglanski & T. E. Higgins (Eds.), *Social psychology: Handbook of basic principles* (2nd ed., pp. 490–515). New York: Guilford Press.
- Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 1–74). New York: Academic Press.
- Florack, A., Scarabis, M., & Bless, H. (2001). When do associations matter? The use of automatic associations toward ethnic groups in person judgments. *Journal of Experimental Social Psychology*, 37, 518–524.
- Ford, T. E., & Kruglanski, A. W. (1995). Effects of epistemic motivations on the use of accessible constructs in social judgment. *Personality and Social Psychology Bulletin*, 21, 950–962.
- Forgas, J. P. (Ed.). (2001). *Handbook of affect and social cognition*. Mahwah, NJ: Erlbaum.
- Förster, J., Higgins, E. T., & Bianco, A. T. (2003). Speed/accuracy decisions in task performance: Built in trade-off of separate strategic concerns. *Organization Behavior and Human Decision Processes*, 90, 148–164.
- Freitas, A. L., & Higgins, E. T. (2002). Enjoying goal-directed action: The role of regulatory fit. *Psychological Science*, 13, 1–6.
- Friedman, R., & Förster, J. (2001). The effects of promotion and prevention cues on creativity. *Journal of Personality and Social Psychology*, 81, 1001–1013.
- Fritzsche, I., Jonas, E., & Fankhänel, T. (2008). The role of control motivation in mortality salience effects on ingroup support and defense. *Journal of Personality and Social Psychology*, 95, 524–541.
- Fu, H., Morris, M. W., Lee, S., Chao, M., Chiu, C., & Hong, Y. (2007). Epistemic motives and cultural conformity: Need for closure, culture, and context as determinants of conflict judgments. *Journal of Personality and Social Psychology*, 92, 191–207.
- Giner-Sorolla, R., & Chaiken, S. (1997). Selective use of heuristic and systematic processing under defense motivation. *Personality and Social Psychology Bulletin*, 23, 84–97.
- Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, 135, 555–588.
- Heine, S. J., Proulx, T., & Vohs, K. D. (2006). The meaning maintenance model: On the coherence of human motivations. *Personality and Social Psychology Review*, 10, 88–110.
- Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 133–168). New York: Guilford Press.
- Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist*, 52, 1280–1300.
- Higgins, E. T. (2000). Making a good decision: Value from fit. *American Psychologist*, 55, 1217–1230.
- Higgins, E. T. (2001). Promotion and prevention experiences: Relating emotions to nonemotional motivational states. In J. P. Forgas (Ed.), *Handbook of affect and social cognition* (pp. 186–211). Mahwah, NJ: Erlbaum.
- Higgins, E. T. (2008). Regulatory fit. In J. Y. Shah & W. L. Gardner (Eds.), *Handbook of motivation science* (pp. 356–372). New York: Guilford Press.
- Higgins, E. T., Idson, L. C., Freitas, A. L., Spiegel, S., & Molden, D. C. (2003). Transfer of value from fit. *Journal of Personality and Social Psychology*, 84, 1140–1153.
- Higgins, E. T., Kruglanski, A. W., & Pierro, A. (2003). Regulatory mode: Locomotion and assessment as distinct orientations. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (pp. 293–344). New York: Academic Press.
- Higgins, E. T., & Molden, D. C. (2003). How strategies for making judgments and decisions affect cognition: Motivated cognition revisited. In G. V. Bodenhausen & A. J. Lambert (Eds.), *Foundations of social cognition: A festschrift in honor of Robert S. Wyer, Jr.* (pp. 211–236). Mahwah, NJ: Erlbaum.
- Higgins, E. T., Roney, C., Crowe, E., & Hymes, C. (1994). Ideal versus ought predilections for approach and avoidance: Distinct self-regulatory systems. *Journal of Personality and Social Psychology*, 66, 276–286.
- Higgins, E. T., & Tykocinski, O. (1992). Self-discrepancies and biographical memory: Personality and cognition at the level of psychological situations. *Personality and Social Psychology Bulletin*, 18, 527–535.
- Hirschberger, G. (2006). Terror management and attributions of blame to innocent victims: Reconciling compassionate and defensive responses. *Journal of Personality and Social Psychology*, 91, 832–844.
- Hong, Y., Morris, M., Chiu, C., & Benet-Martínez, V. (2000). Multicultural minds: A dynamic constructivist approach to culture and cognition. *American Psychologist*, 55, 709–720.
- Hsee, C. K. (1995). Elastic justification: How tempting but task-irrelevant factors influence decisions. *Organizational Behavior and Human Decision Processes*, 62, 330–337.
- Hugenberg, K., Miller, J., & Claypool, H. M. (2007). Categorization and individuation in the cross-race recognition deficit: Toward a solution to an insidious problem. *Journal of Experimental Social Psychology*, 43, 334–340.

- Jonas, E., & Frey, D. (2003). Information search and presentation in advisor-client inter-actions. *Organizational Behavior and Human Decision Processes*, 91, 154–168.
- Jonas, E., Schulz-Hardt, S., & Frey, D. (2005). Giving advice or making decisions in someone else's place: The influence of impression, defense, and accuracy motivation on the search for new information. *Personality and Social Psychology Bulletin*, 31, 977–990.
- Josephs, R. A., & Hahn, E. D. (1995). Bias and accuracy in estimates of task duration. *Organizational Behavior and Human Decision Processes*, 61, 202–213.
- Kay, A. C., Gaucher, D., Napier, J. L., Callan, M. J., & Laurin, K. (2008). God and the government: Testing a compensatory control mechanism for the support of external systems. *Journal of Personality and Social Psychology*, 95, 18–35.
- Kay, A. C., Gaucher, D., Peach, J. M., Friesen, J., Laurin, K., Zanna, M. P., & Spencer, S. J. (2009). Inequality, discrimination, and the power of the status quo: Direct evidence for a motivation to view what is as what should be. *Journal of Personality and Social Psychology*, 97, 421–434.
- Kay, A. C., Jost, J. T., & Young, S. (2005). Victim-derogation and victim-enhancement as alternate routes to system-justification. *Psychological Science*, 16, 240–246.
- Kelley, H. H. (1973). The process of causal attribution. *American Psychologist*, 28, 107–128.
- Klein, W. M., & Kunda, Z. (1992). Motivated person perception: Constructing justifications for desired beliefs. *Journal of Experimental Social Psychology*, 28, 145–168.
- Knowles, M. L., & Gardner, W. L. (2008). Benefits of membership: The activation and amplification of group identities in response to social rejection. *Personality and Social Psychology Bulletin*, 34, 1200–1213.
- Knowles, M. L., Lucas, G. M., Molden, D. C., Gardner, W. L., & Dean, K. K. (2010). There's no substitute for belonging: Self-affirmation following social and non-social threats. *Personality and Social Psychology Bulletin*, 36, 173–186.
- Koenig, A. M., Cesario, J., Molden, D. C., Kosloff, S., & Higgins, E. T. (2009). Incidental experiences of regulatory fit and the processing of persuasive appeals. *Personality and Social Psychology Bulletin*, 35, 1342–1355.
- Kosic, A., Kruglanski, A. W., Pierro, A., & Mannetti, L. (2004). The social cognition of immigrants' acculturation: Effects of the need for closure and the reference group at entry. *Journal of Personality and Social Psychology*, 86, 796–813.
- Kossowska, M. (2007). The role of cognitive inhibition in motivation toward closure. *Personality and Individual Differences*, 42, 1117–1126.
- Kraus, M. W., Piff, P. K., & Keltner, D. (2009). Social class, sense of control, and social explanation. *Journal of Personality and Social Psychology*, 97, 992–1004.
- Kruglanski, A. W. (1996). Motivated social cognition: Principles of the interface. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 493–520). New York: Guilford.
- Kruglanski, A. W., & Freund, T. (1983). The freezing and unfreezing of lay inferences: Effects on impression primacy, ethnic stereotyping and numerical anchoring. *Journal of Experimental Social Psychology*, 19, 448–468.
- Kruglanski, A. W., Orehek, E., Higgins, E. T., Pierro, A., & Shaley, I. (2010). Modes of self-regulation: Assessment and locomotion as independent determinants in goal-pursuit. In R. Hoyle (Ed.), *Handbook of personality and self-regulation* (pp. 375–402). Hoboken, NJ: Wiley.
- Kruglanski, A. W., Pierro, A., Mannetti, L., & De Grada, E. (2006). Groups as epistemic providers: Need for closure and the unfolding of group-centrism. *Psychological Review*, 113, 84–100.
- Kruglanski, A. W., Thompson, E. P., Higgins, E. T., Atash, M. N., Pierro, A., Shah, J. Y., & Spiegel, S. (2000). To "do the right thing" or to "just do it": Locomotion and assessment as distinct self-regulatory imperatives. *Journal of Personality and Social Psychology*, 79, 793–815.
- Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: "Seizing" and "freezing." *Psychological Review*, 103, 263–283.
- Kruglanski, A. W., Webster, D. M., & Klem, A. (1993). Motivated resistance and openness to persuasion in the presence of absence of prior information. *Journal of Personality and Social Psychology*, 65, 861–876.
- Krull, D. S., & Erickson, D. J. (1995). Inferential hopscotch: How people draw social inferences from behavior. *Current Directions in Psychological Science*, 4, 35–38.
- Kunda, Z. (1987). Motivated inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology*, 53, 636–647.
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108, 480–498.
- Kunda, Z., & Santioso, R. (1989). Motivated change in the self-concept. *Journal of Experimental Social Psychology*, 25, 272–285.
- Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, 125, 255–275.
- Lewin, K. (1935). *A dynamic theory of personality*. New York: McGraw-Hill.
- Liberman, N., Molden, D. C., Idson, L. C., & Higgins, E. T. (2001). Promotion and prevention focus on alternative hypotheses: Implications for attributional functions. *Journal of Personality and Social Psychology*, 80, 5–18.
- Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarizations: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37, 2098–2109.
- Lun, J., Sinclair, S., Whitchurch, E. R., & Glenn, C. (2007). (Why) do I think what you think? Epistemic social tuning and implicit prejudice. *Journal of Personality and Social Psychology*, 93, 957–972.
- Maddox, W. T., & Markman, A. B. (2010). The motivation-cognition interface in learning and decision making. *Current Directions in Psychological Science*, 19(2), 106–110.
- Maheswaran, D., & Chaiken, S. (1991). Promoting systematic processing in low-motivation settings: Effect of incongruent information on processing and judgment. *Journal of Personality and Social Psychology*, 61, 13–25.
- Maio, G. R., Pakizeh, A., Cheung, W., & Rees, K. J. (2009). Changing, priming, and acting on values: Effects via motivational relations in a circular model. *Journal of Personality and Social Psychology*, 97, 699–715.
- Maner, J. K. (2009). Anxiety: Proximate processes and ultimate functions. *Social and Personality Psychology Compass*, 3, 798–811.
- Maner, J. K., DeWall, C. N., Baumeister, R. F., & Schaller, M. (2007). Does social exclusion motivate interpersonal reconnection? Resolving the "porcupine problem." *Journal of Personality and Social Psychology*, 92, 42–55.
- Maner, J. K., Gailliot, M. T., Rouby, D. A., & Miller, S. L. (2007). Can't take my eyes off you: Attentional adhesion to

- mates and rivals. *Journal of Personality and Social Psychology*, 93, 389–401.
- Maner, J. K., Kenrick, D. T., Neuberg, S. L., Becker, D. V., Robertson, T., Hofer, B.,...Schaller, M. (2005). Functional projection: How fundamental social motives can bias interpersonal perception. *Journal of Personality and Social Psychology*, 88, 63–78.
- Maner, J. K., Miller, S. L., Rouby, D. A., & Gailliot, M. T. (2009). Intrasexual vigilance: The implicit cognition of romantic rivalry. *Journal of Personality and Social Psychology*, 97, 74–87.
- Mannetti, L., Pierro, A., & Kruglanski, A. W. (2007). Who regrets more after choosing a non-status quo option? Post decisional regret under need for cognitive closure. *Journal of Economic Psychology*, 28, 186–196.
- Mannetti, L., Leder, S., Insalata, L., Pierro, A., Higgins, E. T., & Kruglanski, A. (2009). Priming the ant or the grasshopper in people's mind: How regulatory mode affects intertemporal choices. *European Journal of Social Psychology*, 39, 1120–1125.
- Martin, L., & Clore, G. C. (Eds.). (2001). *Theories of mood and cognition*. Mahwah, NJ: Erlbaum.
- Mauro, R., Pierro, A., Mannetti, L., Higgins, E. T., & Kruglanski, A. W. (2009). The perfect mix: Regulatory complementarity and the speed-accuracy balance in group performance. *Psychological Science*, 20, 681–685.
- Mezulis, A., Abramson, L., Hyde, J. S., & Hankin, B. L. (2004). Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias. *Psychological Bulletin*, 130, 711–746.
- McDonald, H. E., & Hirt, E. R. (1997). When expectancy meets desire: Motivational effects in reconstructive memory. *Journal of Personality and Social Psychology*, 72, 5–23.
- Mishra, S., & Lalumière, M. L. (2010). You can't always get what you want: The motivational effect of need on risky decision-making. *Journal of Experimental Social Psychology*, 46, 605–611.
- Molden, D. C., & Higgins, E. T. (2004). Categorization under uncertainty: Resolving vagueness and ambiguity with eager versus vigilant strategies. *Social Cognition*, 22, 248–277.
- Molden, D. C., & Higgins, E. T. (2005). Motivated thinking. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 295–320). New York: Cambridge University Press.
- Molden, D. C., & Higgins, E. T. (2008). How preferences for eager versus vigilant judgment strategies affect self-serving conclusions. *Journal of Experimental Social Psychology*, 44, 1219–1228.
- Molden, D. C., Lee, A. Y., & Higgins, E. T. (2008). Motivations for promotion and prevention. In J. Shah & W. Gardner (Eds.), *Handbook of motivation science* (pp. 169–187). New York: Guilford.
- Molden, D. C., Lucas, G. M., Gardner, W. L., Dean, K., & Knowles, M. (2009). Motivations for prevention or promotion following social exclusion: Being rejected versus being ignored. *Journal of Personality and Social Psychology*, 96, 415–431.
- Molden, D. C., & Maner, J. K. (in press). How and when exclusion motivates reconnection. In C. N. DeWall (Ed.), *The Oxford handbook of social exclusion*. New York: Oxford University Press.
- Moskowitz, G. B. (1993). Individual differences in social categorization: The influence of personal need for structure on spontaneous trait inferences. *Journal of Personality and Social Psychology*, 65, 132–142.
- Moskowitz, G. B., Gollwitzer, P. M., Wasel, W., & Schaal, B. (1999). Preconscious control of stereotype activation through chronic egalitarian goals. *Journal of Personality and Social Psychology*, 77, 167–184.
- Murray, S. L. (1999). The quest for conviction: Motivated cognition in romantic relationships. *Psychological Inquiry*, 10, 23–34.
- Murray, H. A. (1938). *Explorations in personality*. New York: Oxford University Press.
- Neuberg, S. L. (1989). The goal of forming accurate impressions during social interactions: Attenuating the impact of negative expectancies. *Journal of Personality and Social Psychology*, 56, 374–386.
- Neuberg, S. L., & Fiske, S. T. (1987). Motivational influences on impression formation: Dependency, accuracy-driven attention, and individuating information. *Journal of Personality and Social Psychology*, 53, 431–444.
- Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice Hall.
- Pelham, B. W., & Neter, E. (1995). The effect of motivation on judgment depends on the difficulty of the judgment. *Journal of Personality and Social Psychology*, 68, 581–594.
- Pendry, L. F., & Macrae, C. N. (1994). Stereotypes and mental life: The case of the motivated but thwarted tactician. *Journal of Experimental Social Psychology*, 30, 303–325.
- Petty, R. E., & Wegener, D. T. (1999). The elaboration likelihood model: Current status and controversies. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 41–72). New York: Guilford Press.
- Pierro, A., & Kruglanski, A. W. (2008). "Seizing and freezing" on a significant-person schema: Need for closure and the transference effect in social judgment. *Personality and Social Psychology Bulletin*, 34, 1492–1503.
- Pierro, A., Kruglanski, A. W., & Higgins, E. T. (2006). Regulatory mode and the joys of doing: Effects of "locomotion" and "assessment" on intrinsic and extrinsic task-motivation. *European Journal of Personality*, 20, 355–375.
- Pierro, A., Leder, S., Mannetti, L., Higgins, E. T., Kruglanski, A. W., & Aiello, A. (2008). Regulatory mode effects on counterfactual thinking and regret. *Journal of Experimental Social Psychology*, 44, 321–329.
- Pierro, A., Presaghi, F., Higgins, E. T., & Kruglanski, A. W. (2009). Regulatory mode preferences for autonomy supporting versus controlling instructional styles. *The British Journal of Educational Psychology*, 79, 599–615.
- Pittman, T. S., & D'Agostino, P. R. (1989). Motivation and cognition: Control deprivation and the nature of subsequent information processing. *Journal of Experimental Social Psychology*, 25, 465–480.
- Plaks, J. E., & Stecher, K. (2007). Unexpected improvement, decline, and stasis: A prediction confidence perspective on achievement success and failure. *Journal of Personality and Social Psychology*, 93, 667–684.
- Proulx, T., & Heine, S. J. (2009). Connections from Kafka: Exposure to schema threats improves implicit learning of an artificial grammar. *Psychological Science*, 20, 1125–1131.
- Raymond, J. E., & O'Brien, J. L. (2009). Selective visual attention and motivation: The consequences of value learning in an attentional blink task. *Psychological Science*, 20, 981–988.
- Roese, N. (1997). Counterfactual thinking. *Psychological Bulletin*, 121, 133–148.

- Roese, N. J., Hur, T., & Pennington, G. L. (1999). Counterfactual thinking and regulatory focus: Implications for action versus inaction and sufficiency versus necessity. *Journal of Personality and Social Psychology*, 77, 1109–1120.
- Roets, A., & Van Hiel, A. (2008). Why some hate to dillydally and others do not: The arousal-invoking capacity of decision-making for low and high-scoring need for closure individuals. *Social Cognition*, 26, 333–346.
- Roets, A., Van Hiel, A., Cornelis, I., & Soetens, B. (2008). Determinants of task performance and invested effort: A need for closure by ability interaction analysis. *Personality and Social Psychology Bulletin*, 34, 779–792.
- Sanbonmatsu, D. M., & Fazio, R. H. (1990). The role of attitudes in memory-based decision making. *Journal of Personality and Social Psychology*, 59, 614–622.
- Santioso, R., Kunda, Z., & Fong, G. T. (1990). Motivated recruitment of autobiographical memories. *Journal of Personality and Social Psychology*, 59, 229–241.
- Scholer, A. A., Stroessner, S. J., & Higgins, E. T. (2008). Responding to negativity: How a risky tactic can serve a vigilant strategy. *Journal of Experimental Social Psychology*, 44, 767–774.
- Scholer, A. A., Zou, X., Fujita, K., Stroessner, S. J., & Higgins, E. T. (2010). When risk seeking becomes a motivational necessity. *Journal of Personality and Social Psychology*, 99, 215–231.
- Schuette, R. A., & Fazio, R. H. (1995). Attitude accessibility and motivation as determinants of biased processing: A test of the MODE model. *Personality and Social Psychology Bulletin*, 21, 704–710.
- Schwarz, N., & Clore, G. L. (2007). Feelings and phenomenal experiences. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (2nd ed., pp. 385–407). New York: Guilford.
- Seligman, M. E. P. (1976). *Learned helplessness and depression in animals and men*. Morristown, NJ: General Learning Press.
- Sherman, J. W., Stroessner, S. J., Conrey, F. R., & Azam, O. (2005). Prejudice and stereotype maintenance processes: Attention, attribution, and individuation. *Journal of Personality and Social Psychology*, 89, 607–622.
- Simpson, J. A., Rholes, W. S., & Winterheld, H. A. (2010). Attachment working models twist memories of relationship events. *Psychological Science*, 21, 252–259.
- Sinclair, L., & Kunda, Z. (1999). Reactions to a black professional: Motivated inhibition and activation and conflicting stereotypes. *Journal of Personality and Social Psychology*, 77, 885–904.
- Sinclair, S., Lowery, B. S., Hardin, C. D., & Colangelo, A. (2005). Social tuning of automatic racial attitudes: The role of affiliative motivation. *Journal of Personality and Social Psychology*, 89, 583–592.
- Skitka, L. J., Bauman, C. W., & Lytle, B. L. (2009). Limits on legitimacy: Moral and religious convictions as constraints on deference to authority. *Journal of Personality and Social Psychology*, 97, 567–578.
- Skitka, L. J., Mullen, E., Griffin, T., Hutchinson, S., & Chamberlin, B. (2002). Dispositions, ideological scripts, or motivated correction? Understanding ideological differences in attributions for social problems. *Journal of Personality and Social Psychology*, 83, 470–487.
- Slotter, E. B., & Gardner, W. L. (2009). Where do "you" end and "I" begin? Pre-emptive self-other inclusion as a motivated process. *Journal of Personality and Social Psychology*, 96, 1137–1151.
- Slrull, T. K., Lichtenstein, M., & Rothbart, M. (1985). Associative storage and retrieval processes in person memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11, 316–345.
- Stillwell, A. M., & Baumeister, R. F. (1997). The construction of victim and perpetrator memories: Accuracy and distortion in role-based accounts. *Personality and Social Psychology Bulletin*, 23, 1157–1172.
- Strachman, A., & Gable, S. L. (2006). What you want (and don't want) affects what you see (and don't see): Avoidance social goals and social events. *Personality and Social Psychology Bulletin*, 32, 1446–1458.
- Tetlock, P. E. (1983). Accountability and complexity of thought. *Journal of Personality and Social Psychology*, 45, 74–83.
- Tetlock, P. E. (1985). Accountability: A social check on the fundamental attribution error. *Social Psychology Quarterly*, 48, 227–236.
- Tetlock, P. (1998). Close-call counterfactuals and belief-system defense: I was not almost wrong but I was almost right. *Journal of Personality and Social Psychology*, 75, 639–652.
- Tetlock, P. E., & Boettger, R. (1989). Accountability: A social magnifier of the dilution effect. *Journal of Personality and Social Psychology*, 57, 388–398.
- Tetlock, P. E., & Levi, A. (1982). Attribution bias: On the inconclusiveness of the cognition-motivation debate. *Journal of Experimental Social Psychology*, 18, 68–88.
- Thompson, E. P., Roman, R. J., Moskowitz, G. B., Chaiken, S., & Bargh, J. A. (1994). Systematic processing and the debasing of covert primacy effects in impression formation: Unshackling the motivated perceiver from constraints of accessibility. *Journal of Personality and Social Psychology*, 66, 474–489.
- Tyler, T. R., & Blader, S. L. (2000). *Cooperation in groups: Procedural justice, social identity, and behavioral engagement*. Philadelphia, PA: Psychology Press.
- Vaughn, L. A., O'Rourke, T., Schwartz, S., Malik, J., Petkova, Z., & Trudeau, L. (2006). When two wrongs can make a right: Regulatory nonfit, bias, and correction of judgments. *Journal of Experimental Social Psychology*, 42, 654–661.
- Verplanken, B. (1993). Need for cognition and external information search: Responses to time pressure during decision-making. *Journal of Research in Personality*, 27, 238–252.
- Webster, D. M. (1993). Motivated augmentation and reduction of the over-attribution bias. *Journal of Personality and Social Psychology*, 65, 261–271.
- Warburton, W. A., Williams, K. D., & Cairns, D. R. (2006). When ostracism leads to aggression: The moderating effects of control deprivation. *Journal of Experimental Social Psychology*, 42, 213–220.
- Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J. H., & Cacioppo, J. T. (2010). Making sense by making sentient: Effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology*, 99, 410–435.
- Whitson, J. A., & Galinsky, A. D. (2008). Lacking control increases illusory pattern perception. *Science*, 322, 115–117.
- Zelaznik, H. N., Mone, S., McCabe, G. P., & Thaman, C. (1988). Role of temporal and spatial precision in determining the nature of the speed-accuracy trade-off in aimed-hand movement. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 221–230.

This page intentionally left blank

PART
4

Problem Solving, Intelligence, and Creative Thinking

This page intentionally left blank

Problem Solving

Miriam Bassok and Laura R. Novick

Abstract

This chapter follows the historical development of research on problem solving. It begins with a description of two research traditions that addressed different aspects of the problem-solving process: (1) research on *problem representation* (the Gestalt legacy) that examined how people understand the problem at hand, and (2) research on *search in a problem space* (the legacy of Newell and Simon) that examined how people generate the problem's solution. It then describes some developments in the field that fueled the integration of these two lines of research: work on problem isomorphs, on expertise in specific knowledge domains (e.g., chess, mathematics), and on insight solutions. Next, it presents examples of recent work on problem solving in science and mathematics that highlight the impact of visual perception and background knowledge on how people represent problems and search for problem solutions. The final section considers possible directions for future research.

Key Words: problem solving, representation, search, expertise, insight, Gestalt, diagrams, math, science

Overview

People are confronted with problems on a daily basis, be it trying to extract a broken light bulb from a socket, finding a detour when the regular route is blocked, fixing dinner for unexpected guests, dealing with a medical emergency, or deciding what house to buy. Obviously, the problems people encounter differ in many ways, and their solutions require different types of knowledge and skills. Yet we have a sense that all the situations we classify as problems share a common core. Karl Duncker defined this core as follows: “A problem arises when a living creature has a goal but does not know how this goal is to be reached. Whenever one cannot go from the given situation to the desired situation simply by action [i.e., by the performance of obvious operations], then there has to be recourse to thinking” (Duncker, 1945, p. 1). Consider the broken light bulb. The obvious operation—holding the glass part of the bulb with one’s fingers while unscrewing the

base from the socket—is prevented by the fact that the glass is broken. Thus, there must be “recourse to thinking” about possible ways to solve the problem. For example, one might try mounting half a potato on the broken bulb (we do not know the source of this creative solution, which is described on many “how to” Web sites).

The above definition and examples make it clear that what constitutes a problem for one person may not be a problem for another person, or for that same person at another point in time. For example, the second time one has to remove a broken light bulb from a socket, the solution likely can be retrieved from memory; there is no problem. Similarly, tying shoes may be considered a problem for 5-year-olds but not for readers of this chapter. And, of course, people may change their goal and either no longer have a problem (e.g., take the guests to a restaurant instead of fixing dinner) or attempt to solve a different problem (e.g., decide what restaurant to

go to). Given the highly subjective nature of what constitutes a problem, researchers who study problem solving have often presented people with novel problems that they should be capable of solving and attempted to find regularities in the resulting problem-solving behavior. Despite the variety of possible problem situations, researchers have identified important regularities in the thinking processes by which people (*a*) *represent*, or understand, problem situations and (*b*) *search* for possible ways to get to their goal.

A problem representation is a model constructed by the solver that summarizes his or her understanding of the problem components—the *initial state* (e.g., a broken light bulb in a socket), the *goal state* (the light bulb extracted), and the set of possible *operators* one may apply to get from the initial state to the goal state (e.g., use pliers). According to Reitman (1965), problem components differ in the extent to which they are *well defined*. Some components leave little room for interpretation (e.g., the initial state in the broken light bulb example is relatively well defined), whereas other components may be *ill defined* and have to be defined by the solver (e.g., the possible actions one may take to extract the broken bulb). The solver's representation of the problem guides the search for a possible solution (e.g., possible attempts at extracting the light bulb). This search may, in turn, change the representation of the problem (e.g., finding that the goal cannot be achieved using pliers) and lead to a new search.

Such a recursive process of representation and search continues until the problem is solved or until the solver decides to abort the goal.

Duncker (1945, pp. 28–37) documented the interplay between representation and search based on his careful analysis of one person's solution to the “Radiation Problem” (later to be used extensively in research on analogy, see Holyoak, Chapter 13). This problem requires using some rays to destroy a patient's stomach tumor without harming the patient. At sufficiently high intensity, the rays will destroy the tumor. However, at that intensity, they will also destroy the healthy tissue surrounding the tumor. At lower intensity, the rays will not harm the healthy tissue, but they also will not destroy the tumor. Duncker's analysis revealed that the solver's solution attempts were guided by three distinct problem representations. He depicted these solution attempts as an inverted search tree in which the three main branches correspond to the three general problem representations (Duncker, 1945, p. 32). We reproduce this diagram in Figure 21.1. The desired solution appears on the rightmost branch of the tree, within the general problem representation in which the solver aims to “lower the intensity of the rays on their way through healthy tissue.” The actual solution is to project multiple low-intensity rays at the tumor from several points around the patient “by use of lens.” The low-intensity rays will converge on the tumor, where their individual intensities will sum to a level sufficient to destroy the tumor.

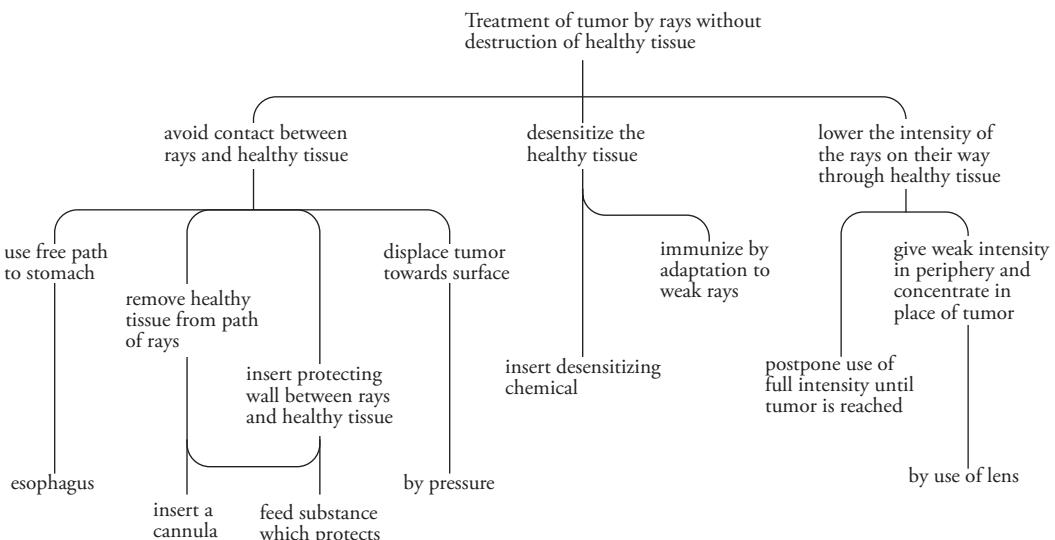


Fig. 21.1 A search-tree representation of one subject's solution to the radiation problem, reproduced from Duncker (1945, p. 32).

Although there are inherent interactions between representation and search, some researchers focus their efforts on understanding the factors that affect how solvers represent problems, whereas others look for regularities in how they search for a solution within a particular representation. Based on their main focus of interest, researchers devise or select problems with solutions that mainly require either constructing a particular representation or finding the appropriate sequence of steps leading from the initial state to the goal state. In most cases, researchers who are interested in problem representation select problems in which one or more of the components are ill defined, whereas those who are interested in search select problems in which the components are well defined. The following examples illustrate, respectively, these two problem types.

The Bird-and-Trains problem (Posner, 1973, pp. 150–151) is a mathematical word problem that tends to elicit two distinct problem representations (see Fig. 21.2a and b):

Two train stations are 50 miles apart. At 2 p.m. one Saturday afternoon two trains start toward each other, one from each station. Just as the trains pull out of the stations, a bird springs into the air in front of the first train and flies ahead to the front of the second train. When the bird reaches the second train, it turns back and flies toward the first train. The bird continues to do this until the trains meet. If both trains travel at the rate of 25 miles per hour and the

bird flies at 100 miles per hour, how many miles will the bird have flown before the trains meet?

Some solvers focus on the back-and-forth path of the bird (Fig. 21.2a). This representation yields a problem that would be difficult for most people to solve (e.g., a series of differential equations). Other solvers focus on the paths of the trains (Fig. 21.2b), a representation that yields a relatively easy distance-rate-time problem.

The Tower of Hanoi problem falls on the other end of the representation-search continuum. It leaves little room for differences in problem representations, and the primary work is to discover a solution path (or the best solution path) from the *initial state* to the *goal state*.

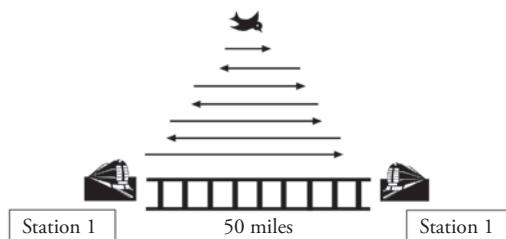
There are three pegs mounted on a base. On the leftmost peg, there are three disks of differing sizes. The disks are arranged in order of size with the largest disk on the bottom and the smallest disk on the top. The disks may be moved one at a time, but only the top disk on a peg may be moved, and at no time may a larger disk be placed on a smaller disk. The goal is to move the three-disk tower from the leftmost peg to the rightmost peg.

Figure 21.3 shows all the possible legal arrangements of disks on pegs. The arrows indicate transitions between states that result from moving a single disk, with the thicker gray arrows indicating the shortest path that connects the initial state to the goal state.

The division of labor between research on representation versus search has distinct historical antecedents and research traditions. In the next two sections, we review the main findings from these two historical traditions. Then, we describe some developments in the field that fueled the integration of these lines of research—work on problem isomorphs, on expertise in specific knowledge domains (e.g., chess, mathematics), and on insight solutions. In the fifth section, we present some examples of recent work on problem solving in science and mathematics. This work highlights the role of visual perception and background knowledge in the way people represent problems and search for problem solutions. In the final section, we consider possible directions for future research.

Our review is by no means an exhaustive one. It follows the historical development of the field and highlights findings that pertain to a wide variety of problems. Research pertaining to specific types

(a) A representation focused on the bird.



(b) A representation focused on the trains.

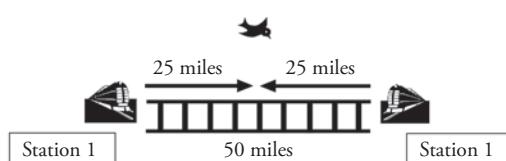


Fig. 21.2 Alternative representations of Posner's (1973) trains-and-bird problem. Adapted from Novick and Hmelo (1994).

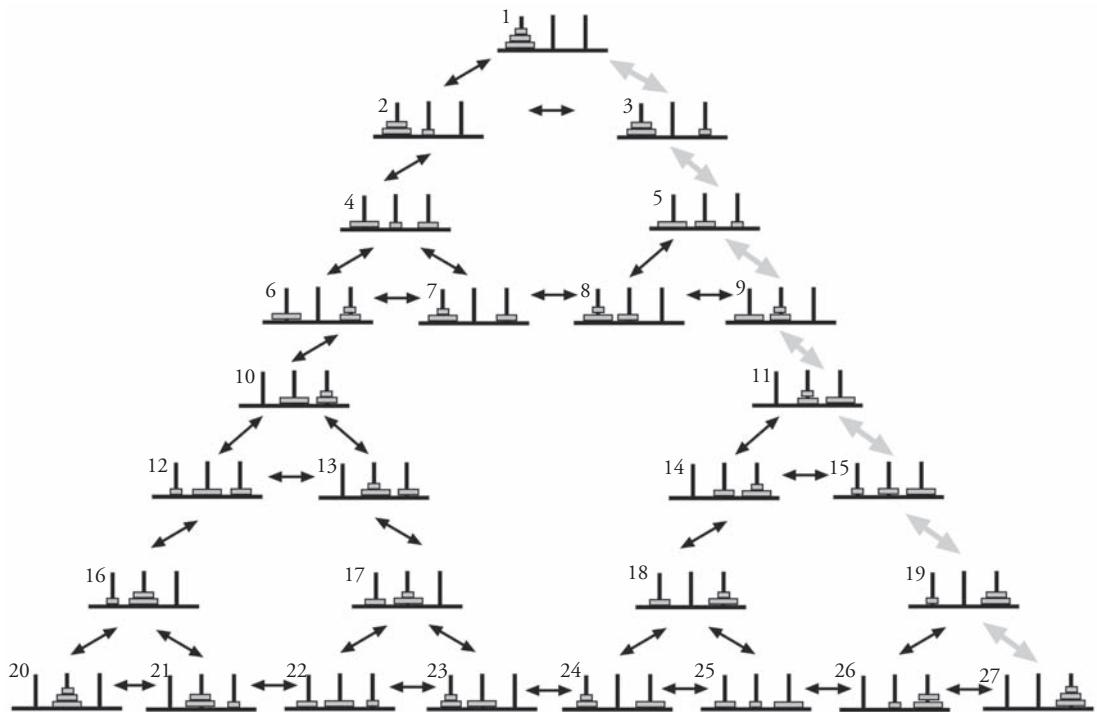


Fig 21.3 All possible problem states for the three-disk Tower of Hanoi problem. The thicker gray arrows show the optimum solution path connecting the initial state (State #1) to the goal state (State #27).

of problems (e.g., medical problems), specific processes that are involved in problem solving (e.g., analogical inferences), and developmental changes in problem solving due to learning and maturation may be found elsewhere in this volume (e.g., Holyoak, Chapter 13; Smith & Ward, Chapter 23; van Steenburgh et al., Chapter 24; Simonton, Chapter 25; Opfer & Siegler, Chapter 30; Hegarty & Stull, Chapter 31; Dunbar & Klahr, Chapter 35; Patel et al., Chapter 37; Lowenstein, Chapter 38; Koedinger & Roll, Chapter 40).

Problem Representation: The Gestalt Legacy

Research on problem representation has its origins in Gestalt psychology, an influential approach in European psychology during the first half of the 20th century. (Behaviorism was the dominant perspective in American psychology at this time.) Karl Duncker published a book on the topic in his native German in 1935, which was translated into English and published 10 years later as the monograph *On Problem-Solving* (Duncker, 1945). Max Wertheimer also published a book on the topic in 1945, titled *Productive Thinking*. An enlarged edition published

posthumously includes previously unpublished material (Wertheimer, 1959). Interestingly, 1945 seems to have been a watershed year for problem solving, as mathematician George Polya's book, *How to Solve It*, also appeared then (a second edition was published 12 years later; Polya, 1957).

The Gestalt psychologists extended the organizational principles of visual perception to the domain of problem solving. They showed that various visual aspects of the problem, as well as the solver's prior knowledge, affect how people understand problems and, therefore, generate problem solutions. The principles of visual perception (e.g., proximity, closure, grouping, good continuation) are directly relevant to problem solving when the physical layout of the problem, or a diagram that accompanies the problem description, elicits inferences that solvers include in their problem representations. Such effects are nicely illustrated by Maier's (1930) nine-dot problem: Nine dots are arrayed in a 3x3 grid, and the task is to connect all the dots by drawing four straight lines without lifting one's pencil from the paper. People have difficulty solving this problem because their initial representations generally include a constraint, inferred from the configuration

of the dots, that the lines should not go outside the boundary of the imaginary square formed by the outer dots. With this constraint, the problem cannot be solved (but see Adams, 1979). Without this constraint, the problem may be solved as shown in Figure 21.4 (though the problem is still difficult for many people; see Weisberg & Alba, 1981).

The nine-dot problem is a classic *insight* problem (see van Steenburgh et al., Chapter 24). According to the Gestalt view (e.g., Duncker, 1945; Kohler, 1925; Maier, 1931; see Ohlsson, 1984, for a review), the solution to an insight problem appears suddenly, accompanied by an “aha!” sensation, immediately following the sudden “restructuring” of one’s understanding of the problem (i.e., a change in the problem representation): “The decisive points in thought-processes, the moments of sudden comprehension, of the ‘Aha!’, of the new, are always at the same time moments in which such a sudden restructuring of the thought-material takes place” (Duncker, 1945, p. 29). For the nine-dot problem, one view of the required restructuring is that the solver relaxes the constraint implied by the perceptual form of the problem and realizes that the lines may, in fact, extend past the boundary of the imaginary square. Later in the chapter, we present more recent accounts of insight.

The entities that appear in a problem also tend to evoke various inferences that people incorporate into their problem representations. A classic demonstration of this is the phenomenon of *functional fixedness*, introduced by Duncker (1945): If an object is habitually used for a certain purpose (e.g., a box serves as a container), it is difficult to see

that object as having properties that would enable it to be used for a dissimilar purpose. Duncker’s basic experimental paradigm involved two conditions that varied in terms of whether the object that was crucial for solution was initially used for a function other than that required for solution.

Consider the candles problem—the best known of the five “practical problems” Duncker (1945) investigated. Three candles are to be mounted at eye height on a door. On the table, for use in completing this task, are some tacks and three boxes. The solution is to tack the three boxes to the door to serve as platforms for the candles. In the control condition, the three boxes were presented to subjects empty. In the functional-fixedness condition, they were filled with candles, tacks, and matches. Thus, in the latter condition, the boxes initially served the function of container, whereas the solution requires that they serve the function of platform. The results showed that 100% of the subjects who received empty boxes solved the candles problem, compared with only 43% of subjects who received filled boxes. Every one of the five problems in this study showed a difference favoring the control condition over the functional-fixedness condition, with average solution rates across the five problems of 97% and 58%, respectively.

The function of the objects in a problem can be also “fixed” by their most recent use. For example, Birch and Rabinowitz (1951) had subjects perform two consecutive tasks. In the first task, people had to use either a switch or a relay to form an electric circuit. After completing this task, both groups of subjects were asked to solve Maier’s (1931) two-ropes problem. The solution to this problem requires tying an object to one of the ropes and making the rope swing as a pendulum. Subjects could create the pendulum using either the object from the electric-circuit task or the other object. Birch and Rabinowitz found that subjects avoided using the same object for two unrelated functions. That is, those who used the switch in the first task made the pendulum using the relay, and vice versa. The explanations subjects subsequently gave for their object choices revealed that they were unaware of the functional-fixedness constraint they imposed on themselves.

In addition to investigating people’s solutions to such practical problems as irradiating a tumor, mounting candles on the wall, or tying ropes, the Gestalt psychologists examined how people understand and solve mathematical problems that require domain-specific knowledge. For example,

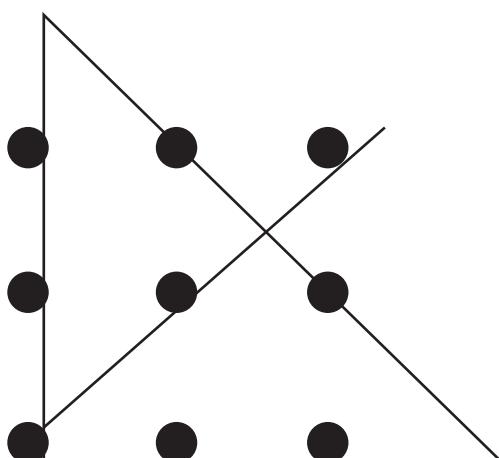


Fig. 21.4 A solution to the nine-dot problem.

Wertheimer (1959) observed individual differences in students' learning and subsequent application of the formula for finding the area of a parallelogram (see Fig. 21.5a). Some students understood the logic underlying the learned formula (i.e., the fact that a parallelogram can be transformed into a rectangle by cutting off a triangle from one side and pasting it onto the other side) and exhibited "productive thinking"—using the same logic to find the area of the quadrilateral in Figure 21.5b and the irregularly shaped geometric figure in Figure 21.5c. Other students memorized the formula and exhibited "reproductive thinking"—reproducing the learned solution only to novel parallelograms that were highly similar to the original one.

The psychological study of human problem solving faded into the background after the demise of the Gestalt tradition (during World War II), and problem solving was investigated only sporadically until Allen Newell and Herbert Simon's (1972) landmark book *Human Problem Solving* sparked a flurry of research on this topic. Newell and Simon adopted and refined Duncker's (1945) methodology of collecting and analyzing the think-aloud protocols that accompany problem solutions and extended Duncker's conceptualization of a problem solution as a search tree. However, their initial work did not aim to extend the Gestalt findings

pertaining to problem representation. Instead, as we explain in the next section, their objective was to identify the general-purpose strategies people use in searching for a problem solution.

Search in a Problem Space: The Legacy of Newell and Simon

Newell and Simon (1972) wrote a magnum opus detailing their theory of problem solving and the supporting research they conducted with various collaborators. This theory was grounded in the information-processing approach to cognitive psychology and guided by an analogy between human and artificial intelligence (i.e., both people and computers being "Physical Symbol Systems," Newell & Simon, 1976; see Doumas & Hummel, Chapter 5). They conceptualized problem solving as a process of search through a *problem space* for a path that connects the initial state to the goal state—a metaphor that alludes to the visual or spatial nature of problem solving (Simon, 1990). The term *problem space* refers to the solver's representation of the task as presented (Simon, 1978). It consists of (1) a set of knowledge states (the initial state, the goal state, and all possible intermediate states), (2) a set of operators that allow movement from one knowledge state to another, (3) a set of constraints, and (4) local information about the path one is taking through the space (e.g., the current knowledge state and how one got there).

We illustrate the components of a problem space for the three-disk Tower of Hanoi problem, as depicted in Figure 21.3. The initial state appears at the top (State #1) and the goal state at the bottom right (State #27). The remaining knowledge states in the figure are possible intermediate states. The current knowledge state is the one at which the solver is located at any given point in the solution process. For example, the current state for a solver who has made three moves along the optimum solution path would be State #9. The solver presumably would know that he or she arrived at this state from State #5. This knowledge allows the solver to recognize a move that involves backtracking. The three operators in this problem are moving each of the three disks from one peg to another. These operators are subject to the constraint that a larger disk may not be placed on a smaller disk.

Newell and Simon (1972), as well as other contemporaneous researchers (e.g., Atwood & Polson, 1976; Greene, 1974; Thomas, 1974), examined how people traverse the spaces of various well-defined problems (e.g., the Tower of Hanoi,

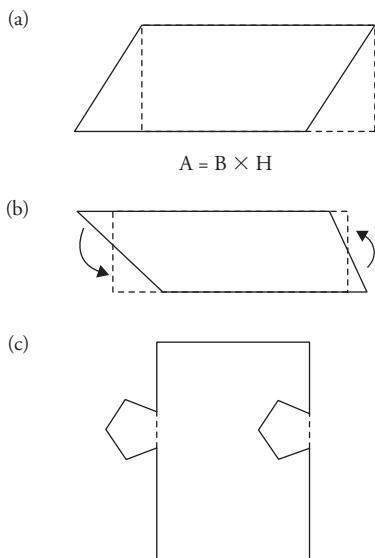


Fig. 21.5 Finding the area of (a) a parallelogram, (b) a quadrilateral, and (c) an irregularly shaped geometric figure. The solid lines indicate the geometric figures whose areas are desired. The dashed lines show how to convert the given figures into rectangles (i.e., they show solutions with understanding).

Hobbits and Orcs). They discovered that solvers' search is guided by a number of shortcut strategies, or *heuristics*, which are likely to get the solver to the goal state without an extensive amount of search. Heuristics are often contrasted with *algorithms*—methods that are guaranteed to yield the correct solution. For example, one could try every possible move in the three-disk Tower of Hanoi problem and, eventually, find the correct solution. Although such an exhaustive search is a valid algorithm for this problem, for many problems its application is very time consuming and impractical (e.g., consider the game of chess).

In their attempts to identify people's search heuristics, Newell and Simon (1972) relied on two primary methodologies: think-aloud protocols and computer simulations. Their use of think-aloud protocols brought a high degree of scientific rigor to the methodology used by Duncker (1945; see Ericsson & Simon, 1980). Solvers were required to say out loud everything they were thinking as they solved the problem, that is, everything that went through their verbal working memory. Subjects' verbalizations—their think-aloud protocols—were tape-recorded and then transcribed verbatim for analysis. This method is extremely time consuming (e.g., a transcript of one person's solution to the cryptarithmetic problem DONALD + GERALD = ROBERT, with D = 5, generated a 17-page transcript), but it provides a detailed record of the solver's ongoing solution process.

An important caveat to keep in mind while interpreting a subject's verbalizations is that "a protocol is relatively reliable only for what it positively contains, but not for that which it omits" (Duncker, 1945, p. 11). Ericsson and Simon (1980) provided an in-depth discussion of the conditions under which this method is valid (but see Russo, Johnson, & Stephens, 1989, for an alternative perspective). To test their interpretation of a subject's problem solution, inferred from the subject's verbal protocol, Newell and Simon (1972) created a computer simulation program and examined whether it solved the problem the same way the subject did. To the extent that the computer simulation provided a close approximation of the solver's step-by-step solution process, it lent credence to the researcher's interpretation of the verbal protocol.

Newell and Simon's (1972) most famous simulation was the General Problem Solver or GPS (Ernst & Newell, 1969). GPS successfully modeled human solutions to problems as different as the Tower of

Hanoi and the construction of logic proofs using a single general-purpose heuristic: *means-ends analysis*. This heuristic captures people's tendency to devise a solution plan by setting *subgoals* that could help them achieve their final goal. It consists of the following steps: (1) Identify a difference between the current state and the goal (or *subgoal*) state; (2) Find an operator that will remove (or reduce) the difference; (3a) If the operator can be directly applied, do so, or (3b) If the operator cannot be directly applied, set a subgoal to remove the obstacle that is preventing execution of the desired operator; (4) Repeat steps 1–3 until the problem is solved. Next, we illustrate the implementation of this heuristic for the Tower of Hanoi problem, using the problem space in Figure 21.3.

As can be seen in Figure 21.3, a key difference between the initial state and the goal state is that the large disk is on the wrong peg (step 1). To remove this difference (step 2), one needs to apply the operator "move-large-disk." However, this operator cannot be applied because of the presence of the medium and small disks on top of the large disk. Therefore, the solver may set a subgoal to move that two-disk tower to the middle peg (step 3b), leaving the rightmost peg free for the large disk. A key difference between the initial state and this new subgoal state is that the medium disk is on the wrong peg. Because application of the move-medium-disk operator is blocked, the solver sets another subgoal to move the small disk to the right peg. This subgoal can be satisfied immediately by applying the move-small-disk operator (step 3a), generating State #3. The solver then returns to the previous subgoal—moving the tower consisting of the small and medium disks to the middle peg. The differences between the current state (#3) and the subgoal state (#9) can be removed by first applying the move-medium-disk operator (yielding State #5) and then the move-small-disk operator (yielding State #9). Finally, the move-large-disk operator is no longer blocked. Hence, the solver moves the large disk to the right peg, yielding State #11.

Notice that the subgoals are stacked up in the order in which they are generated, so that they pop up in the order of last in first out. Given the first subgoal in our example, repeated application of the means-ends analysis heuristic will yield the shortest-path solution, indicated by the large gray arrows. In general, subgoals provide direction to the search and allow solvers to plan several moves ahead. By assessing progress toward a required subgoal rather

than the final goal, solvers may be able to make moves that otherwise seem unwise. To take a concrete example, consider the transition from State #1 to State #3 in Figure 21.3. Comparing the initial state to the goal state, this move seems unwise because it places the small disk on the bottom of the right peg, whereas it ultimately needs to be at the top of the tower on that peg. But comparing the initial state to the solver-generated subgoal state of having the medium disk on the middle peg, this is exactly where the small disk needs to go.

Means-ends analysis and various other heuristics (e.g., the *hill-climbing* heuristic that exploits the similarity, or distance, between the state generated by the next operator and the goal state; *working backward* from the goal state to the initial state) are flexible strategies that people often use to successfully solve a large variety of problems. However, the generality of these heuristics comes at a cost: They are relatively weak and fallible (e.g., in the means-ends solution to the problem of fixing a hole in a bucket, “Dear Liza” leads “Dear Henry” in a loop that ends back at the initial state; the lyrics of this famous song can be readily found on the Web). Hence, although people use general-purpose heuristics when they encounter novel problems, they replace them as soon as they acquire experience with and sufficient knowledge about the particular problem space (e.g., Anzai & Simon, 1979).

Despite the fruitfulness of this research agenda, it soon became evident that a fundamental weakness was that it minimized the importance of people’s background knowledge. Of course, Newell and Simon (1972) were aware that problem solutions require relevant knowledge (e.g., the rules of logical proofs, or rules for stacking disks). Hence, in programming GPS, they supplemented every problem they modeled with the necessary background knowledge. This practice highlighted the generality and flexibility of means-ends analysis but failed to capture how people’s background knowledge affects their solutions. As we discussed in the previous section, domain knowledge is likely to affect how people represent problems and, therefore, how they generate problem solutions. Moreover, as people gain experience solving problems in a particular knowledge domain (e.g., math, physics), they change their representations of these problems (e.g., Chi, Feltovich, & Glaser, 1981; Haverty, Koedinger, Klahr, & Alibali, 2000; Schoenfeld & Herrmann, 1982) and learn domain-specific heuristics (e.g., Polya, 1957; Schoenfeld, 1979) that trump the general-purpose strategies.

It is perhaps inevitable that the two traditions in problem-solving research—one emphasizing representation and the other emphasizing search strategies—would eventually come together. In the next section we review developments that led to this integration.

The Two Legacies Converge

Because Newell and Simon (1972) aimed to discover the strategies people use in searching for a solution, they investigated problems that minimized the impact of factors that tend to evoke differences in problem representations, of the sort documented by the Gestalt psychologists. In subsequent work, however, Simon and his collaborators showed that such factors are highly relevant to people’s solutions of well-defined problems, and Simon (1986) incorporated these findings into the theoretical framework that views problem solving as search in a problem space.

In this section, we first describe illustrative examples of this work. We then describe research on insight solutions that incorporates ideas from the two legacies described in the previous sections.

Relevance of the Gestalt Ideas to the Solution of Search Problems

In this subsection we describe two lines of research by Simon and his colleagues, and by other researchers, that document the importance of perception and of background knowledge to the way people search for a problem solution. The first line of research used variants of relatively well-defined riddle problems that had the same structure (i.e., “problem isomorphs”) and, therefore, supposedly the same problem space. It documented that people’s search depended on various perceptual and conceptual inferences they tended to draw from a specific instantiation of the problem’s structure. The second line of research documented that people’s search strategies crucially depend on their domain knowledge and on their prior experience with related problems.

PROBLEM ISOMORPHS

Hayes and Simon (1977) used two variants of the Tower of Hanoi problem that, instead of disks and pegs, involved monsters and globes that differed in size (small, medium, and large). In both variants, the initial state had the small monster holding the large globe, the medium-sized monster holding the small globe, and the large monster holding the medium-sized globe. Moreover, in both variants the

goal was for each monster to hold a globe proportionate to its own size. The only difference between the problems concerned the description of the operators. In one variant (“transfer”), subjects were told that the monsters could transfer the globes from one to another as long as they followed a set of rules, adapted from the rules in the original Tower of Hanoi problem (e.g., only one globe may be transferred at a time). In the other variant (“change”), subjects were told that the monsters could shrink and expand themselves according to a set of rules, which corresponded to the rules in the transfer version of the problem (e.g., only one monster may change its size at a time). Despite the isomorphism of the two variants, subjects conducted their search in two qualitatively different problem spaces, which led to solution times for the change variant being almost twice as long as those for the transfer variant. This difference arose because subjects could more readily envision and track an object that was changing its location with every move than one that was changing its size.

Recent work by Patsenko and Altmann (2010) found that, even in the standard Tower of Hanoi problem, people’s solutions involve object-bound routines that depend on perception and selective attention. The subjects in their study solved various Tower of Hanoi problems on a computer. During the solution of a particular “critical” problem, the computer screen changed at various points without subjects’ awareness (e.g., a disk was added, such that a subject who started with a five-disc tower ended with a six-disc tower). Patsenko and Altmann found that subjects’ moves were guided by the configurations of the objects on the screen rather than by solution plans they had stored in memory (e.g., the next subgoal).

The Gestalt psychologists highlighted the role of perceptual factors in the formation of problem representations (e.g., Maier’s, 1930, nine-dot problem) but were generally silent about the corresponding implications for how the problem was solved (although they did note effects on solution accuracy). An important contribution of the work on people’s solutions of the Tower of Hanoi problem and its variants was to show the relevance of perceptual factors to the application of various operators during search for a problem solution—that is, to the *how* of problem solving. In the next section, we describe recent work that documents the involvement of perceptual factors in how people understand and use equations and diagrams in the context of solving math and science problems.

Kotovsky, Hayes, and Simon (1985) further investigated factors that affect people’s representation and search in isomorphs of the Tower of Hanoi problem. In one of their isomorphs, three disks were stacked on top of each other to form an inverted pyramid, with the smallest disc on the bottom and the largest on top. Subjects’ solutions of the inverted pyramid version were similar to their solutions of the standard version that has the largest disc on the bottom and the smallest on top. However, the two versions were solved very differently when subjects were told that the discs represent acrobats. Subjects readily solved the version in which they had to place a small acrobat on the shoulders of a large one, but they refrained from letting a large acrobat stand on the shoulders of a small one. In other words, object-based inferences that draw on people’s semantic knowledge affected the solution of search problems, much as they affect the solution of the ill-defined problems investigated by the Gestalt psychologists (e.g., Duncker’s, 1945, candles problem). In the next section, we describe more recent work that shows similar effects in people’s solutions to mathematical word problems.

The work on differences in the representation and solution of problem isomorphs is highly relevant to research on analogical problem solving (or analogical transfer), which examines when and how people realize that two problems that differ in their cover stories have a similar structure (or a similar problem space) and, therefore, can be solved in a similar way. This research shows that minor differences between example problems, such as the use of X-rays versus ultrasound waves to fuse a broken filament of a light bulb, can elicit different problem representations that significantly affect the likelihood of subsequent transfer to novel problem analogs (Holyoak & Koh, 1987). Analogical transfer has played a central role in research on human problem solving, in part because it can shed light on people’s understanding of a given problem and its solution and in part because it is believed to provide a window onto understanding and investigating creativity (see Smith & Ward, Chapter 23). We briefly mention some findings from the analogy literature in the next subsection on expertise, but we do not discuss analogical transfer in detail because this topic is covered elsewhere in this volume (Holyoak, Chapter 13).

EXPERTISE AND ITS DEVELOPMENT

In another line of research, Simon and his colleagues examined how people solve ecologically valid

problems from various rule-governed and knowledge-rich domains. They found that people's level of expertise in such domains, be it in chess (Chase & Simon, 1973; Gobet & Simon, 1996), mathematics (Hinsley, Hayes, & Simon, 1977; Paige & Simon, 1966), or physics (Larkin, McDermott, Simon, & Simon, 1980; Simon & Simon, 1978), plays a crucial role in how they represent problems and search for solutions. This work, and the work of numerous other researchers, led to the discovery (and rediscovery, see Duncker, 1945) of important differences between experts and novices, and between "good" and "poor" students.

One difference between experts and novices pertains to pattern recognition. Experts' attention is quickly captured by familiar configurations within a problem situation (e.g., a familiar configuration of pieces in a chess game). In contrast, novices' attention is focused on isolated components of the problem (e.g., individual chess pieces). This difference, which has been found in numerous domains, indicates that experts have stored in memory many meaningful groups (chunks) of information: for example, chess (Chase & Simon, 1973), circuit diagrams (Egan & Schwartz, 1979), computer programs (McKeithen, Reitman, Rueter, & Hirtle, 1981), medicine (Coughlin & Patel, 1987; Myles-Worsley, Johnston, & Simons, 1988), basketball and field hockey (Allard & Starkes, 1991), and figure skating (Deakin & Allard, 1991).

The perceptual configurations that domain experts readily recognize are associated with stored solution plans and/or compiled procedures (Anderson, 1982). As a result, experts' solutions are much faster than, and often qualitatively different from, the piece-meal solutions that novice solvers tend to construct (e.g., Larkin et al., 1980). In effect, experts often see the solutions that novices have yet to compute (e.g., Chase & Simon, 1973; Novick & Sherman, 2003, 2008). These findings have led to the design of various successful instructional interventions (e.g., Catrambone, 1998; Kellman et al., 2008). For example, Catrambone (1998) perceptually isolated the subgoals of a statistics problem. This perceptual chunking of meaningful components of the problem prompted novice students to self-explain the meaning of the chunks, leading to a conceptual understanding of the learned solution. In the next section, we describe some recent work that shows the beneficial effects of perceptual pattern recognition on the solution of familiar mathematics problems, as well as the potentially detrimental effects

of familiar perceptual chunks to understanding and reasoning with diagrams depicting evolutionary relationships among taxa.

Another difference between experts and novices pertains to their understanding of the solution-relevant problem structure. Experts' knowledge is highly organized around domain principles, and their problem representations tend to reflect this principled understanding. In particular, they can extract the solution-relevant structure of the problems they encounter (e.g., meaningful causal relations among the objects in the problem; see Cheng & Buehner, Chapter 12). In contrast, novices' representations tend to be bound to surface features of the problems that may be irrelevant to solution (e.g., the particular objects in a problem). For example, Chi, Feltovich, and Glaser (1981) examined how students with different levels of physics expertise group mechanics word problems. They found that advanced graduate students grouped the problems based on the physics principles relevant to the problems' solutions (e.g., conservation of energy, Newton's second law). In contrast, undergraduates who had successfully completed an introductory course in mechanics grouped the problems based on the specific objects involved (e.g., pulley problems, inclined plane problems). Other researchers have found similar results in the domains of biology, chemistry, computer programming, and math (Adelson, 1981; Kindfield, 1993/1994; Kozma & Russell, 1997; McKeithen et al., 1981; Silver, 1979, 1981; Weiser & Shertz, 1983).

The level of domain expertise and the corresponding representational differences are, of course, a matter of degree. With increasing expertise, there is a gradual change in people's focus of attention from aspects that are not relevant to solution to those that are (e.g., Deakin & Allard, 1991; Hardiman, Dufresne, & Mestre, 1989; McKeithen et al., 1981; Myles-Worsley et al., 1988; Schoenfeld & Herrmann, 1982; Silver, 1981). Interestingly, Chi, Bassok, Lewis, Reimann, and Glaser (1989) found similar differences in focus on structural versus surface features among a group of novices who studied worked-out examples of mechanics problems. These differences, which echo Wertheimer's (1959) observations of individual differences in students' learning about the area of parallelograms, suggest that individual differences in people's interests and natural abilities may affect whether, or how quickly, they acquire domain expertise.

An important benefit of experts' ability to focus their attention on solution-relevant aspects of problems is that they are more likely than novices to recognize analogous problems that involve different objects and cover stories (e.g., Chi et al., 1989; Novick, 1988; Novick & Holyoak, 1991; Wertheimer, 1959) or that come from other knowledge domains (e.g., Bassok & Holyoak, 1989; Dunbar, 2001; Goldstone & Sakamoto, 2003). For example, Bassok and Holyoak (1989) found that, after learning to solve arithmetic-progression problems in algebra, subjects spontaneously applied these algebraic solutions to analogous physics problems that dealt with constantly accelerated motion. Note, however, that experts and good students do not simply ignore the surface features of problems. Rather, as was the case in the problem isomorphs we described earlier (Kotovsky et al., 1985), they tend to use such features to infer what the problem's structure could be (e.g., Alibali, Bassok, Solomon, Syc, & Goldin-Meadow, 1999; Blessing & Ross, 1996). For example, Hinsley et al. (1977) found that, after reading no more than the first few words of an algebra word problem, expert solvers classified the problem into a likely problem category (e.g., a work problem, a distance problem) and could predict what questions they might be asked and the equations they likely would need to use.

Surface-based problem categorization has a heuristic value (Medin & Ross, 1989): It does not ensure a correct categorization (Blessing & Ross, 1996), but it does allow solvers to retrieve potentially appropriate solutions from memory and to use them, possibly with some adaptation, to solve a variety of novel problems. Indeed, although experts exploit surface-structure correlations to save cognitive effort, they have the capability to realize that a particular surface cue is misleading (Hegarty, Mayer, & Green, 1992; Lewis & Mayer, 1987; Martin & Bassok, 2005; Novick 1988, 1995; Novick & Holyoak, 1991). It is not surprising, therefore, that experts may revert to novice-like heuristic methods when solving problems under pressure (e.g., Beilock, 2008) or in sub-domains in which they have general but not specific expertise (e.g., Patel, Groen, & Arocha, 1990).

Relevance of Search to Insight Solutions

We introduced the notion of insight in our discussion of the nine-dot problem in the section on the Gestalt tradition. The Gestalt view (e.g., Duncker, 1945; Maier, 1931; see Ohlsson, 1984, for a review) was that insight problem solving is characterized

by an initial work period during which no progress toward solution is made (i.e., an impasse), a sudden restructuring of one's problem representation to a more suitable form, followed immediately by the sudden appearance of the solution. Thus, solving problems by insight was believed to be all about representation, with essentially no role for a step-by-step solution process (i.e., search). Subsequent and contemporary researchers have generally concurred with the Gestalt view that getting the right representation is crucial. However, research has shown that insight solutions do not necessarily arise suddenly or full blown after restructuring (e.g., Weisberg & Alba, 1981); and even when they do, the underlying solution process (in this case outside of awareness) may reflect incremental progress toward the goal (Bowden & Jung-Beeman, 2003; Durso, Rea, & Dayton, 1994; Novick & Sherman, 2003).

"Demystifying insight," to borrow a phrase from Bowden, Jung-Beeman, Fleck, and Kounios (2005), requires explaining (1) why solvers initially reach an impasse in solving a problem for which they have the necessary knowledge to generate the solution, (2) how the restructuring occurred, and (3) how it led to the solution. A detailed discussion of these topics appears elsewhere in this volume (van Steenburgh et al., Chapter 24). Here, we describe briefly three recent theories that have attempted to account for various aspects of these phenomena: Knoblich, Ohlsson, Haider, and Rhenius's (1999) representational change theory, MacGregor, Ormerod, and Chronicle's (2001) progress monitoring theory, and Bowden et al.'s (2005) neurological model. We then propose the need for an integrated approach to demystifying insight that considers both representation and search.

According to Knoblich et al.'s (1999) representational change theory, problems that are solved with insight are highly likely to evoke initial representations in which solvers place inappropriate constraints on their solution attempts, leading to an impasse. An impasse can be resolved by revising one's representation of the problem. Knoblich and his colleagues tested this theory using Roman numeral matchstick arithmetic problems in which solvers must move one stick to a new location to change a false numerical statement (e.g., $I = II + II$) into a statement that is true. According to representational change theory, re-representation may occur through either constraint relaxation or chunk decomposition. (The solution to the example problem is to change $II +$ to $III -$, which requires both

methods of re-representation, yielding $I = III - II$). Good support for this theory has been found based on measures of solution rate, solution time, and eye fixation (Knoblich et al., 1999; Knoblich, Ohlsson, & Raney, 2001; Öllinger, Jones, & Knoblich, 2008).

Progress monitoring theory (MacGregor et al., 2001) was proposed to account for subjects' difficulty in solving the nine-dot problem, which has traditionally been classified as an insight problem. According to this theory, solvers use the hill-climbing search heuristic to solve this problem, just as they do for traditional search problems (e.g., Hobbits and Orcs). In particular, solvers are hypothesized to monitor their progress toward solution using a criterion generated from the problem's current state. If solvers reach criterion failure, they seek alternative solutions by trying to relax one or more problem constraints. MacGregor et al. found support for this theory using several variants of the nine-dot problem (also see Ormerod, MacGregor, & Chronicle, 2002). Jones (2003) suggested that progress monitoring theory provides an account of the solution process up to the point an impasse is reached and representational change is sought, at which point representational change theory picks up and explains how insight may be achieved. Hence, it appears that a complete account of insight may require an integration of concepts from the Gestalt (representation) and Newell and Simon's (search) legacies.

Bowden et al.'s (2005) neurological model emphasizes the overlap between problem solving and language comprehension, and it hinges on differential processing in the right and left hemispheres. They proposed that an impasse is reached because initial processing of the problem produces strong activation of information irrelevant to solution in the left hemisphere. At the same time, weak semantic activation of alternative semantic interpretations, critical for solution, occurs in the right hemisphere. Insight arises when the weakly activated concepts reinforce each other, eventually rising above the threshold required for conscious awareness. Several studies of problem solving using compound remote associates problems, involving both behavioral and neuroimaging data, have found support for this model (Bowden & Jung-Beeman, 1998, 2003; Jung-Beeman & Bowden, 2000; Jung-Beeman et al., 2004; also see Moss, Kotovsky, & Cagan, 2011).

Note that these three views of insight have received support using three quite distinct types of

problems (Roman numeral matchstick arithmetic problems, the nine-dot problem, and compound remote associates problems, respectively). It remains to be established, therefore, whether these accounts can be generalized across problems. Kershaw and Ohlsson (2004) argued that insight problems are difficult because the key behavior required for solution may be hindered by perceptual factors (the Gestalt view), background knowledge (so expertise may be important; e.g., see Novick & Sherman, 2003, 2008), and/or process factors (e.g., those affecting search). From this perspective, solving visual problems (e.g., the nine-dot problem) with insight may call upon more general visual processes, whereas solving verbal problems (e.g., anagrams, compound remote associates) with insight may call upon general verbal/semantic processes.

Summary

The work we reviewed in this section shows the relevance of problem representation (the Gestalt legacy) to the way people search the problem space (the legacy of Newell and Simon), and the relevance of search to the solution of insight problems that require a representational change. In addition to this inevitable integration of the two legacies, the work we described here underscores the fact that problem solving crucially depends on perceptual factors and on the solvers' background knowledge. In the next section, we describe some recent work that shows the involvement of these factors in the solution of problems in math and science.

Effects of Perception and Knowledge in Problem Solving in Academic Disciplines

Although the use of puzzle problems continues in research on problem solving, especially in investigations of insight, many contemporary researchers tackle problem solving in knowledge-rich domains, often in academic disciplines (e.g., mathematics, biology, physics, chemistry, meteorology). In this section, we provide a sampling of this research that highlights the importance of visual perception and background knowledge for successful problem solving.

The Role of Visual Perception

We stated at the outset that a problem representation (e.g., the problem space) is a model of the problem constructed by solvers to summarize their understanding of the problem's essential nature. This informal definition refers to the *internal*

representations people construct and hold in working memory. Of course, people may also construct various *external representations* (Markman, 1999) and even manipulate those representations to aid in solution (see Hegarty & Stull, Chapter 31). For example, solvers often use paper and pencil to write notes or draw diagrams, especially when solving problems from formal domains (e.g., Cox, 1999; Kindfield, 1993/1994; S. Schwartz, 1971). In problems that provide solvers with external representation, such as the Tower of Hanoi problem, people's planning and memory of the current state is guided by the actual configurations of disks on pegs (Garber & Goldin-Meadow, 2002) or by the displays they see on a computer screen (Chen & Holyoak, 2010; Patsenko & Altmann, 2010).

In STEM (science, technology, engineering, and mathematics) disciplines, it is common for problems to be accompanied by diagrams or other external representations (e.g., equations) to be used in determining the solution. Larkin and Simon (1987) examined whether isomorphic sentential and diagrammatic representations are interchangeable in terms of facilitating solution. They argued that although the two formats may be equivalent in the sense that all of the information in each format can be inferred from the other format (informational equivalence), the ease or speed of making inferences from the two formats might differ (lack of computational equivalence). Based on their analysis of several problems in physics and math, Larkin and Simon further argued for the general superiority of diagrammatic representations (but see Mayer & Gallini, 1990, for constraints on this general conclusion).

Novick and Hurley (2001, p. 221) succinctly summarized the reasons for the general superiority of diagrams (especially abstract or schematic diagrams) over verbal representations: They "(a) simplify complex situations by discarding unnecessary details (e.g., Lynch, 1990; Winn, 1989), (b) make abstract concepts more concrete by mapping them onto spatial layouts with familiar interpretational conventions (e.g., Winn, 1989), and (c) substitute easier perceptual inferences for more computationally intensive search processes and sentential deductive inferences (Barwise & Etchemendy, 1991; Larkin & Simon, 1987)." Despite these benefits of diagrammatic representations, there is an important caveat, noted by Larkin and Simon (1987, p. 99) at the very end of their paper: "Although every diagram supports some easy perceptual inferences,

nothing ensures that these inferences must be useful in the problem-solving process." We will see evidence of this in several of the studies reviewed in this section.

Next we describe recent work on perceptual factors that are involved in people's use of two types of external representations that are provided as part of the problem in two STEM disciplines: equations in algebra and diagrams in evolutionary biology. Although we focus here on effects of perceptual factors per se, it is important to note that such factors only influence performance when subjects have background knowledge that supports differential interpretation of the alternative diagrammatic depictions presented (Hegarty, Canham, & Fabricant, 2010).

EQUATIONS

In the previous section, we described the work of Patsenko and Altmann (2010) that shows direct involvement of visual attention and perception in the sequential application of move operators during the solution of the Tower of Hanoi problem. A related body of work documents similar effects in tasks that require the interpretation and use of mathematical equations (Goldstone, Landy, & Son, 2010; Landy & Goldstone, 2007a, b). For example, Landy and Goldstone (2007b) varied the spatial proximity of arguments to the addition (+) and multiplication (*) operators in algebraic equations, such that the spatial layout of the equation was either consistent or inconsistent with the order-of-operations rule that multiplication precedes addition. In *consistent equations*, the space was narrower around multiplication than around addition (e.g., $g*m + r*w = m*g + w*r$), whereas in *inconsistent equations* this relative spacing was reversed (e.g., $s * n + e * c = n * s + c * e$). Subjects' judgments of the validity of such equations (i.e., whether the expressions on the two sides of the equal sign are equivalent) were significantly faster and more accurate for consistent than inconsistent equations.

In discussing these findings and related work with other external representations, Goldstone et al. (2010) proposed that experience with solving domain-specific problems leads people to "rig up" their perceptual system such that it allows them to look at the problem in a way that is consistent with the correct rules. Similar logic guides the Perceptual Learning Modules developed by Kellman and his collaborators to help students interpret and use algebraic equations and graphs (Kellman et al., 2008;

Kellman, Massey, & Son, 2009). These authors argued and showed that, consistent with the previously reviewed work on expertise, perceptual training with particular external representations supports the development of perceptual fluency. This fluency, in turn, supports students' subsequent use of these external representations for problem solving.

This research suggests that extensive experience with particular equations or graphs may lead to perceptual fluency that could replace the more mindful application of domain-specific rules. Fisher, Borchert, and Bassok (2011) reported results from algebraic-modeling tasks that are consistent with this hypothesis. For example, college students were asked to represent verbal statements with algebraic equations, a task that typically elicits systematic errors (e.g., Clement, Lochhead, & Monk, 1981). Fisher et al. found that such errors were very common when subjects were asked to construct "standard form" equations ($y = ax$), which support fluent left-to-right translation of words to equations, but were relatively rare when subjects were asked to construct nonstandard division-format equations ($x = y/a$) that do not afford such translation fluency.

In part because of the left-to-right order in which people process equations, which mirrors the linear order in which they process text, equations have traditionally been viewed as sentential representations. However, Landy and Goldstone (2007a) have proposed that equations also share some properties with diagrammatic displays and that, in fact, in some ways they are processed like diagrams. That is, spatial information is used to represent and to support inferences about syntactic structure. This hypothesis received support from Landy and Goldstone's (2007b) results, described earlier, in which subjects' judgments of the validity of equations were affected by the Gestalt principle of grouping: Subjects did better when the grouping was consistent rather than inconsistent with the underlying structure of the problem (order of operations). Moreover, Landy and Goldstone (2007a) found that when subjects wrote their own equations they grouped numbers and operators (+, *, =) in a way that reflected the hierarchical structure imposed by the order-of-operations rule.

DIAGRAMS

In a recent line of research, Novick and Catley (2007; Novick, Catley, & Funk, 2010; Novick, Shade, & Catley, 2011) have examined effects of

the spatial layout of diagrams depicting the evolutionary history of a set of taxa on people's ability to reason about patterns of relationship among those taxa. We consider here their work that investigates the role of another Gestalt perceptual principle—good continuation—in guiding students' reasoning. According to this principle, a continuous line is perceived as a single entity (Kellman, 2000). Consider the diagrams shown in Figure 21.6. Each is a cladogram, a diagram that depicts nested sets of taxa that are related in terms of levels of most recent common ancestry. For example, chimpanzees and starfish are more closely related to each other than either is to spiders. The supporting evidence for their close relationship is their most recent common ancestor, which evolved the novel character of having radial cleavage. Spiders do not share this ancestor and thus do not have this character.

Cladograms are typically drawn in two isomorphic formats, which Novick and Catley (2007) referred to as trees and ladders. Although these formats are informationally equivalent (Larkin & Simon, 1987), Novick and Catley's (2007) research shows that they are not computationally equivalent (Larkin & Simon, 1987). Imagine that you are given evolutionary relationships in the ladder format, such as in Figure 21.6a (but without the four characters—hydrostatic skeleton, bilateral symmetry, radial cleavage, and trocophore larvae—and associated short lines indicating their locations on the cladogram), and your task is to translate that diagram to the tree format. A correct translation is shown in Figure 21.6b. Novick and Catley (2007) found that college students were much more likely to get such problems correct when the presented cladogram was in the nested circles (e.g., Figure 21.6d) rather than the ladder format. Because the Gestalt principle of good continuation makes the long slanted line at the base of the ladder appear to represent a single hierarchical level, a common translation error for the ladder to tree problems was to draw a diagram such as that shown in Figure 21.6c.

The difficulty that good continuation presents for interpreting relationships depicted in the ladder format extends to answering reasoning questions as well. Novick and Catley (unpublished data) asked comparable questions about relationships depicted in the ladder and tree formats. For example, using the cladograms depicted in Figures 21.6a and 21.6b, consider the following questions: (a) Which taxon—jellyfish or earthworm—is the closest evolutionary relation to starfish, and what evidence

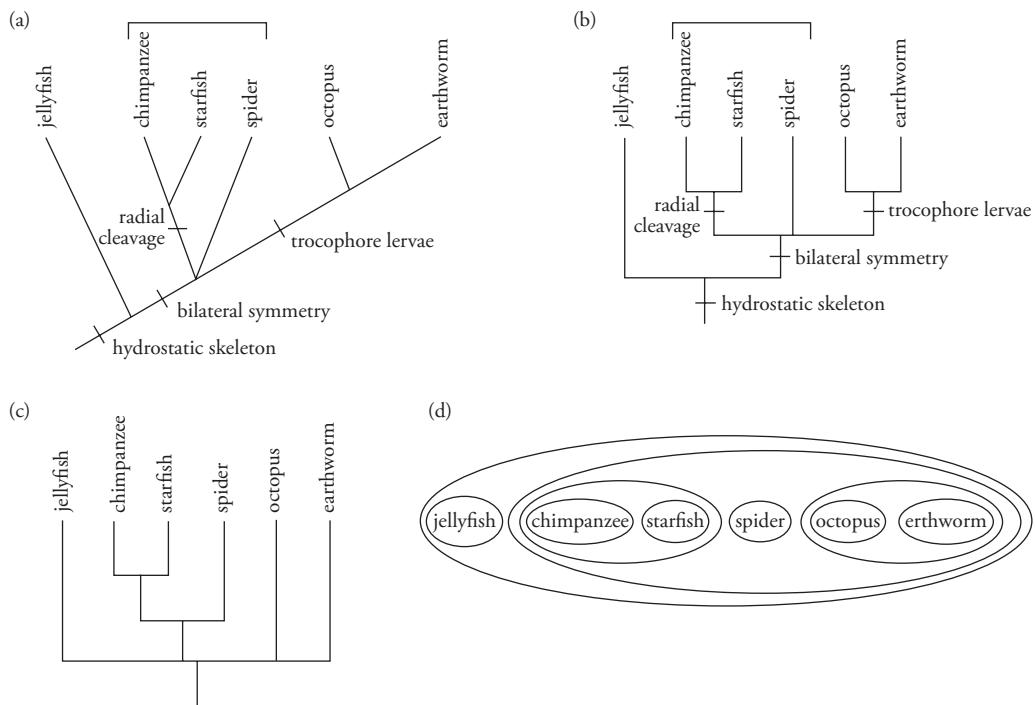


Fig. 21.6 Four cladograms depicting evolutionary relationships among six animal taxa. Cladogram (a) is in the ladder format, cladograms (b) and (c) are in the tree format, and cladogram (d) is in the nested circles format. Cladograms (a), (b), and (d) are isomorphic.

supports your answer? (b) Do the bracketed taxa comprise a clade (a set of taxa consisting of the most recent common ancestor and all of its descendants), and what evidence supports your answer? For both such questions, students had higher accuracy and evidence quality composite scores when the relationships were depicted in the tree than the ladder format.

If the difficulty in extracting the hierarchical structure of the ladder format is due to good continuation (which leads problem solvers to interpret continuous lines that depict multiple hierarchical levels as depicting only a single level), then a manipulation that breaks good continuation at the points where a new hierarchical level occurs should improve understanding. Novick et al. (2010) tested this hypothesis using a translation task by manipulating whether characters that are the markers for the most recent common ancestor of each nested set of taxa were included on the ladders. Figure 21.6a shows a ladder with such characters. As predicted, translation accuracy increased dramatically simply by adding these characters to the ladders, despite the additional information subjects had to account for in their translations.

The Role of Background Knowledge

As we mentioned earlier, the specific entities in the problems people encounter evoke inferences that affect how people represent these problems (e.g., the candle problem; Duncker, 1945) and how they apply the operators in searching for the solution (e.g., the disks vs. acrobats versions of the Tower of Hanoi problem; Kotovsky et al., 1985). Such object-based inferences draw on people's knowledge about the properties of the objects (e.g., a box is a container, an acrobat is a person who can be hurt). Here, we describe the work of Bassok and her colleagues, who found that similar inferences affect how people select mathematical procedures to solve problems in various formal domains. This work shows that the objects in the texts of mathematical word problems affect how people represent the problem situation (i.e., the *situation model* they construct; Kintsch & Greeno, 1985) and, in turn, lead them to select mathematical models that have a corresponding structure. To illustrate, a word problem that describes constant change in the rate at which ice is melting off a glacier evokes a model of continuous change, whereas a word problem that describes constant change in the rate at which ice is

delivered to a restaurant evokes a model of discrete change. These distinct situation models lead subjects to select corresponding visual representations (e.g., Bassok & Olseth, 1995) and solutions methods, such as calculating the average change over time versus adding the consecutive changes (e.g., Alibali et al., 1999).

In a similar manner, people draw on their general knowledge to infer how the objects in a given problem are related to each other and construct mathematical solutions that correspond to these inferred object relations. For example, a word problem that involves doctors from two hospitals elicits a situation model in which the two sets of doctors play symmetric roles (e.g., work with each other), whereas a mathematically isomorphic problem that involves mechanics and cars elicits a situation model in which the sets play asymmetric roles (e.g., mechanics fix cars). The mathematical solutions people construct to such problems reflect this difference in symmetry (Bassok, Wu, & Olseth, 1995). In general, people tend to add objects that belong to the same taxonomic category (e.g., doctors + doctors) but divide functionally related objects (e.g., cars + mechanics). People establish this correspondence by a process of analogical alignment between semantic and arithmetic relations, which Bassok and her colleagues refer to as “semantic alignment” (Bassok, Chase, & Martin, 1998; Doumas, Bassok, Guthormsen, & Hummel, 2006; Fisher, Bassok, & Osterhout, 2010).

Semantic alignment occurs very early in the solution process and can prime arithmetic facts that are potentially relevant to the problem solution (Bassok, Pedigo, & Oskarsson, 2008). Although such alignments can lead to erroneous solutions, they have a high heuristic value because, in most textbook problems, object relations indeed correspond to analogous mathematical relations (Bassok et al., 1998). Interestingly, unlike in the case of reliance on specific surface-structure correlations (e.g., the keyword “more” typically appears in word problems that require addition; Lewis & Mayer, 1987), people are more likely to exploit semantic alignment when they have more, rather than less modeling experience. For example, Martin and Bassok (2005) found very strong semantic-alignment effects when subjects solved simple division word problems, but not when they constructed algebraic equations to represent the relational statements that appeared in the problems. Of course, these subjects had significantly more experience with solving numerical word

problems than with constructing algebraic models of relational statements. In a subsequent study, Fisher and Bassok (2009) found semantic-alignment effects for subjects who constructed correct algebraic models, but not for those who committed modeling errors.

Conclusions and Future Directions

In this chapter, we examined two broad components of the problem-solving process: representation (the Gestalt legacy) and search (the legacy of Newell and Simon). Although many researchers choose to focus their investigation on one or the other of these components, both Duncker (1945) and Simon (1986) underscored the necessity to investigate their interaction, as the representation one constructs for a problem determines (or at least constrains) how one goes about trying to generate a solution, and searching the problem space may lead to a change in problem representation. Indeed, Duncker’s (1945) initial account of one subject’s solution to the radiation problem was followed up by extensive and experimentally sophisticated work by Simon and his colleagues and by other researchers, documenting the involvement of visual perception and background knowledge in how people represent problems and search for problem solutions.

The relevance of perception and background knowledge to problem solving illustrates the fact that, when people attempt to find or devise ways to reach their goals, they draw on a variety of cognitive resources and engage in a host of cognitive activities. According to Duncker (1945), such goal-directed activities may include (a) placing objects into categories and making inferences based on category membership, (b) making inductive inferences from multiple instances, (c) reasoning by analogy, (d) identifying the causes of events, (e) deducing logical implications of given information, (f) making legal judgments, and (g) diagnosing medical conditions from historical and laboratory data. As this list suggests, many of the chapters in the present volume describe research that is highly relevant to the understanding of problem-solving behavior. We believe that important advancements in problem-solving research would emerge by integrating it with research in other areas of thinking and reasoning, and that research in these other areas could be similarly advanced by incorporating the insights gained from research on what has more traditionally been identified as problem solving.

As we have described in this chapter, many of the important findings in the field have been established by a careful investigation of various riddle problems. Although there are good methodological reasons for using such problems, many researchers choose to investigate problem solving using ecologically valid educational materials. This choice, which is increasingly common in contemporary research, provides researchers with the opportunity to apply their basic understanding of problem solving to benefit the design of instruction and, at the same time, allows them to gain a better understanding of the processes by which domain knowledge and educational conventions affect the solution process. We believe that the trend of conducting educationally relevant research is likely to continue, and we expect a significant expansion of research on people's understanding and use of dynamic and technologically rich external representations (e.g., Kellman et al., 2008; Mayer, Griffith, Jurkowitz, & Rothman, 2008; Richland & McDonough, 2010; Son & Goldstone, 2009). Such investigations are likely to yield both practical and theoretical payoffs.

References

- Adams, J. L. (1979). *Conceptual blockbusting: A guide to better ideas* (2nd ed.). New York: Norton.
- Adelson, B. (1981). Problem solving and the development of abstract categories in programming languages. *Memory and Cognition*, 9, 422–433.
- Alibali, M. W., Bassok, M., Solomon, K. O., Syc, S. E., & Goldin-Meadow, S. (1999). Illuminating mental representations through speech and gesture. *Psychological Science*, 10, 327–333.
- Allard, F., & Starkes, J. L. (1991). Motor-skill experts in sports, dance, and other domains. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 126–152). New York: Cambridge University Press.
- Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, 89, 369–406.
- Anzai, Y., & Simon, H. A. (1979). The theory of learning by doing. *Psychological Review*, 86, 124–140.
- Atwood, M. E., & Polson, P. G. (1976). A process model for water jug problems. *Cognitive Psychology*, 8, 191–216.
- Barwise, J., & Etchemendy, J. (1991). Visual information and valid reasoning. In W. Zimmermann & S. Cunningham (Eds.), *Visualization in teaching and learning mathematics* (pp. 9–24). Washington, DC: Mathematical Association of America.
- Bassok, M., Chase, V. M., & Martin, S. A. (1998). Adding apples and oranges: Alignment of semantic and formal knowledge. *Cognitive Psychology*, 35, 99–134.
- Bassok, M., & Holyoak, K. J. (1989). Interdomain transfer between isomorphic topics in algebra and physics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 153–166.
- Bassok, M., & Olseth, K. L. (1995). Object-based representations: Transfer between cases of continuous and discrete models of change. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 1522–1538.
- Bassok, M., Pedigo, S. F., & Oskarsson, A. T. (2008). Priming addition facts with semantic relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34, 343–352.
- Bassok, M., Wu, L., & Olseth, L. K. (1995). Judging a book by its cover: Interpretative effects of content on problem solving transfer. *Memory and Cognition*, 23, 354–367.
- Beilock, S. L. (2008). Math performance in stressful situations. *Current Directions in Psychological Science*, 17, 339–343.
- Birch, H. G., & Rabinowitz, H. S. (1951). The negative effect of previous experience on productive thinking. *Journal of Experimental Psychology*, 41, 122–126.
- Blessing, S. B., & Ross, B. H. (1996). Content effects in problem categorization and problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 792–810.
- Bowden, E. M., & Jung-Beeman, M. (1998). Getting the right idea: Semantic activation in the right hemisphere may help solve insight problems. *Psychological Science*, 6, 435–440.
- Bowden, E. M., & Jung-Beeman, M. (2003). Aha! Insight experience correlates with solution activation in the right hemisphere. *Psychonomic Bulletin and Review*, 10, 730–737.
- Bowden, E. M., Jung-Beeman, M., Fleck, J., & Kounios, J. (2005). New approaches to demystifying insight. *Trends in Cognitive Sciences*, 9, 322–328.
- Catrambone, R. (1998). The subgoal-learning model: Creating better examples so that students can solve novel problems. *Journal of Experimental Psychology: General*, 127, 355–376.
- Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, 4, 55–81.
- Chen, D., & Holyoak, K. J. (2010). Enhancing acquisition of intuition versus planning in problem solving. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1875–1880). Austin, TX: Cognitive Science Society.
- Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13, 145–182.
- Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121–152.
- Clement, J., Lochhead, J., & Monk, G. S. (1981). Translation difficulties in learning mathematics. *The American Mathematical Monthly*, 88, 286–290.
- Coughlin, L. D., & Patel, V. L. (1987). Processing of critical information by physicians and medical students. *Journal of Medical Education*, 62, 818–828.
- Cox, R. (1999). Representation construction, externalized cognition and individual differences. *Learning and Instruction*, 9, 343–363.
- Deakin, J. M., & Allard, F. (1991). Skilled memory in expert figure skaters. *Memory and Cognition*, 19, 79–86.
- Doumas, L. A. A., Bassok, M., Guthormsen, A., & Hummel, J. E. (2006). Theory of reflexive relational generalization. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1246–1250). Mahwah, NJ: Erlbaum.
- Dunbar, K. (2001). The analogical paradox: Why analogy is so easy in naturalistic settings, yet so difficult in the psychological laboratory. In D. Gentner, K. J. Holyoak, &

- B. Kokinov (Eds.), *Analogy: Perspectives from cognitive science* (pp. 313–362). Cambridge, MA: MIT Press.
- Duncker, K. (1945). On problem-solving (L. S. Lees, Trans.). *Psychological Monographs*, 58 (Whole No. 270). (Original work published 1935).
- Durso, F. T., Rea, C. B., & Dayton, T. (1994). Graph-theoretic confirmation of restructuring during insight. *Psychological Science*, 5, 94–98.
- Egan, D. E., & Schwartz, B. J. (1979). Chunking in the recall of symbolic drawings. *Memory and Cognition*, 7, 149–158.
- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, 87, 215–251.
- Ernst, G. W., & Newell, A. (1969). *GPS: A case study in generality and problem solving*. New York: Academic Press.
- Fisher, K. J., & Bassok, M. (2009). Analogical alignments in algebraic modeling. In B. Kokinov, D. Gentner, & K. J. Holyoak (Eds.), *Proceedings of the 2nd International Analogy Conference* (pp. 137–144). Sofia, Bulgaria: New Bulgarian University Press.
- Fisher, K. J., Bassok, M., & Osterhout, L. (2010). When two plus two does not equal four: Event-related potential responses to semantically incongruous arithmetic word problems. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1571–1576). Austin, TX: Cognitive Science Society.
- Fisher, K. J., Borchert, K., & Bassok, M. (2011). Following the standard form: Effects of equation format on algebraic modeling. *Memory and Cognition*, 39, 502–515.
- Garber, P., & Goldin-Meadow, S. (2002). Gesture offers insight into problem solving in adults and children. *Cognitive Science*, 26, 817–831.
- Gobet, F., & Simon, H. (1996). Recall of rapidly presented random chess positions is a function of skill. *Psychonomic Bulletin and Review*, 3, 159–163.
- Goldstone, R. L., Landy, D. H., & Son, J. Y. (2010). The education of perception. *Topics in Cognitive Science*, 2, 265–284.
- Goldstone, R. L., & Sakamoto, J. Y. (2003). The transfer of abstract principles governing complex adaptive systems. *Cognitive Psychology*, 46, 414–466.
- Greeno, J. G. (1974). Hobbits and orcs: Acquisition of a sequential concept. *Cognitive Psychology*, 6, 270–292.
- Hardiman, P. T., Dufresne, R., & Mestre, J. P. (1989). The relation between problem categorization and problem solving among experts and novices. *Memory and Cognition*, 17, 627–638.
- Haverty, L. A., Koedinger, K. R., Klahr, D., & Alibali, M. W. (2000). Solving induction problems in mathematics: Not-so-trivial Pursuit. *Cognitive Science*, 24, 249–298.
- Hayes, J. R., & Simon, H. A. (1977). Psychological differences among problem isomorphs. In N. J. Castellan, D. B. Pisoni, & G. R. Potts (Eds.), *Cognitive theory* (Vol. 2, pp. 21–44). Hillsdale, NJ: Erlbaum.
- Hegarty, M., Canham, M. S., & Fabricant, S. I. (2010). Thinking about the weather: How display salience and knowledge affect performance in a graphic inference task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36, 37–53.
- Hegarty, M., Mayer, R. E., & Green, C. E. (1992). Comprehension of arithmetic word problems: Evidence from students' eye fixations. *Journal of Educational Psychology*, 84, 76–84.
- Hinsley, D. A., Hayes, J. R., & Simon, H. A. (1977). From words to equations: Meaning and representation in algebra word problems. In D. Hinsley, M. Just., & P. Carpenter (Eds.), *Cognitive processes in comprehension* (pp. 89–106). Hillsdale, NJ: Erlbaum.
- Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory and Cognition*, 15, 332–340.
- Jones, G. (2003). Testing two cognitive theories of insight. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 1017–1027.
- Jung-Beeman, M., & Bowden, E. M. (2000). The right hemisphere maintains solution-related activation for yet-to-be solved insight problems. *Memory and Cognition*, 28, 1231–1241.
- Jung-Beeman, M., Bowden, E. M., Haberman, J., Fryniarek, J. L., Arambel-Liu, S., Greenblatt, R., ... Kounios, J. (2004). Neural activity when people solve verbal problems with insight. *PLOS Biology*, 2, 500–510.
- Kellman, P. J. (2000). An update on Gestalt psychology. In B. Landau, J. Sabini, J. Jonides, & E. Newport (Eds.), *Perception, cognition, and language: Essays in honor of Henry and Lila Gleitman* (pp. 157–190). Cambridge, MA: MIT Press.
- Kellman, P. J., Massey, C. M., & Son, J. Y. (2009). Perceptual learning modules in mathematics: Enhancing students' pattern recognition, structure extraction, and fluency. *Topics in Cognitive Science*, 1, 1–21.
- Kellman, P. J., Massey, C., Roth, Z., Burke, T., Zucker, J., Saw, A., ... Wise, J. A. (2008). Perceptual learning and the technology of expertise. *Pragmatics and Cognition*, 16, 356–405.
- Kershaw, T. C., & Ohlsson, S. (2004). Multiple causes of difficulty in insight: The case of the nine-dot problem. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 3–13.
- Kindfield, A. C. H. (1993/1994). Biology diagrams: Tools to think with. *Journal of the Learning Sciences*, 3, 1–36.
- Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review*, 92, 109–129.
- Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 1534–1555.
- Knoblich, G., Ohlsson, S., & Raney, G. E. (2001). An eye movement study of insight problem solving. *Memory and Cognition*, 29, 1000–1009.
- Kohler, W. (1925). *The mentality of apes*. New York: Harcourt Brace.
- Kotovsky, K., Hayes, J. R., & Simon, H. A. (1985). Why are some problems hard? Evidence from Tower of Hanoi. *Cognitive Psychology*, 17, 248–294.
- Kozma, R. B., & Russell, J. (1997). Multimedia and understanding: Expert and novice responses to different representations of chemical phenomena. *Journal of Research in Science Teaching*, 34, 949–968.
- Landy, D., & Goldstone, R. L. (2007a). Formal notations are diagrams: Evidence from a production task. *Memory and Cognition*, 35, 2033–2040.
- Landy, D., & Goldstone, R. L. (2007b). How abstract is symbolic thought? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 720–733.
- Larkin, J. H., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Models of competence in solving physics problems. *Cognitive Science*, 4, 317–345.

- Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, *11*, 65–99.
- Lewis, A. B., & Mayer, R. E. (1987). Students' miscomprehension of relational statements in arithmetic word problems. *Journal of Educational Psychology*, *79*, 363–371.
- Lynch, M. (1990). The externalized retina: Selection and mathematization in the visual documentation of objects in the life sciences. In M. Lynch & S. Woolgar (Eds.), *Representation in scientific practice* (pp. 153–186). Cambridge, MA: MIT Press.
- MacGregor, J. N., Ormerod, T. C., & Chronicle, E. P. (2001). Information processing and insight: A process model of performance on the nine-dot and related problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 176–201.
- Maier, N. (1930). Reasoning in humans. I. On direction. *Journal of Comparative Psychology*, *10*, 15–43.
- Maier, N. (1931). Reasoning in humans. II. The solution of a problem and its appearance in consciousness. *Journal of Comparative Psychology*, *12*, 181–194.
- Markman, A. B. (1999). *Knowledge representation*. Mahwah, NJ: Erlbaum.
- Martin, S. A., & Bassok, M. (2005). Effects of semantic cues on mathematical modeling: Evidence from word-problem solving and equation construction tasks. *Memory and Cognition*, *33*, 471–478.
- Mayer, R. E., & Gallini, J. K. (1990). When is an illustration worth ten thousand words? *Journal of Educational Psychology*, *82*, 715–726.
- Mayer, R. E., Griffith, E., Jurkowitz, I. T. N., & Rothman, D. (2008). Increased interestingness of extraneous details in a multimedia science presentation leads to decreased learning. *Journal of Experimental Psychology: Applied*, *14*, 329–339.
- McKeithen, K. B., Reitman, J. S., Rueter, H. H., & Hirtle, S. C. (1981). Knowledge organization and skill differences in computer programmers. *Cognitive Psychology*, *13*, 307–325.
- Medin, D. L., & Ross, B. H. (1989). The specific character of abstract thought: Categorization, problem solving, and induction. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 5, pp. 189–223). Hillsdale, NJ: Erlbaum.
- Moss, J., Kotovsky, K., & Cagan, J. (2011). The effect of incidental hints when problems are suspended before, during, and after an impasse. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 140–148.
- Myles-Worsley, M., Johnston, W. A., & Simons, M. A. (1988). The influence of expertise on X-ray image processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 553–557.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
- Newell, A., & Simon, H. A. (1976). Computer science as empirical enquiry: Symbols and search. *Communications of the ACM*, *19*, 113–126.
- Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 510–520.
- Novick, L. R. (1995). Some determinants of successful analogical transfer in the solution of algebra word problems. *Thinking and Reasoning*, *1*, 5–30.
- Novick, L. R., & Catley, K. M. (2007). Understanding phylogenies in biology: The influence of a Gestalt perceptual principle. *Journal of Experimental Psychology: Applied*, *13*, 197–223.
- Novick, L. R., Catley, K. M., & Funk, D. J. (2010). Characters are key: The effect of synapomorphies on cladogram comprehension. *Evolution: Education and Outreach*, *3*, 539–547.
- Novick, L. R., & Hmelo, C. E. (1994). Transferring symbolic representations across non-isomorphic problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1296–1321.
- Novick, L. R., & Holyoak, K. J. (1991). Mathematical problem solving by analogy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 398–415.
- Novick, L. R., & Hurley, S. M. (2001). To matrix, network, or hierarchy: That is the question. *Cognitive Psychology*, *42*, 158–216.
- Novick, L. R., Shade, C. K., & Catley, K. M. (2011). Linear versus branching depictions of evolutionary history: Implications for diagram design. *Topics in Cognitive Science*, *3*(3), 536–559.
- Novick, L. R., & Sherman, S. J. (2003). On the nature of insight solutions: Evidence from skill differences in anagram solution. *The Quarterly Journal of Experimental Psychology*, *56A*, 351–382.
- Novick, L. R., & Sherman, S. J. (2008). The effects of superficial and structural information on on-line problem solving for good versus poor anagram solvers. *The Quarterly Journal of Experimental Psychology*, *61*, 1098–1120.
- Ohlsson, S. (1984). Restructuring revisited I. Summary and critique of the Gestalt theory of problem solving. *Scandinavian Journal of Psychology*, *25*, 65–78.
- Öllinger, M., Jones, G., & Knoblich, G. (2008). Investigating the effect of mental set on insight problem solving. *Experimental Psychology*, *55*, 269–282.
- Ormerod, T. C., MacGregor, J. N., & Chronicle, E. P. (2002). Dynamics and constraints in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 791–799.
- Paige, J. M., & Simon, H. A. (1966). Cognitive processes in solving algebra word problems. In B. Kleinmuntz (Ed.), *Problem solving: Research, method, and theory* (pp. 51–119). New York: Wiley.
- Patel, V. L., Groen, G. J., & Arocha, J. F. (1990). Medical expertise as a function of task difficulty. *Memory and Cognition*, *18*, 394–406.
- Patsenko, E. G., & Altmann, E. M. (2010). How planful is routine behavior? A selective attention model of performance in the Tower of Hanoi. *Journal of Experimental Psychology: General*, *139*, 95–116.
- Polya, G. (1957). *How to solve it* (2nd ed.). Princeton, NJ: Princeton University Press.
- Posner, M. I. (1973). *Cognition: An introduction*. Glenview, IL: Scott, Foresman and Company.
- Reitman, W. R. (1965). *Cognition and thought*. New York: Wiley.
- Richland, L. E., & McDonough, I. M. (2010). Learning by analogy: Discriminating between potential analogs. *Contemporary Educational Psychology*, *35*, 28–43.
- Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory and Cognition*, *17*, 759–769.
- Schoenfeld, A. H. (1979). Explicit heuristic training as a variable in problem-solving performance. *Journal for Research in Mathematics Education*, *10*, 173–187.
- Schoenfeld, A. H., & Herrmann, D. J. (1982). Problem perception and knowledge structure in expert and novice mathematical

- problem solvers. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8, 484–494.
- Schwartz, S. H. (1971). Modes of representation and problem solving: Well evolved is half solved. *Journal of Experimental Psychology*, 91, 347–350.
- Silver, E. A. (1979). Student perceptions of relatedness among mathematical verbal problems. *Journal for Research in Mathematics Education*, 10, 195–210.
- Silver, E. A. (1981). Recall of mathematical problem information: Solving related problems. *Journal for Research in Mathematics Education*, 12, 54–64.
- Simon, D. P., & Simon, H. A. (1978). Individual differences in solving physics problems. In R. Siegler (Ed.), *Children's thinking: What develops?* (pp. 325–348). Hillsdale, NJ: Erlbaum.
- Simon, H. A. (1978). Information-processing theory of human problem solving. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes* (Vol. 5, pp. 271–295). Hillsdale, NJ: Erlbaum.
- Simon, H. A. (1986). The information processing explanation of Gestalt Phenomena. *Computers in Human Behavior*, 2, 241–255.
- Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology*, 41, 1–19.
- Son, J. Y., & Goldstone, R. L. (2009). Fostering general transfer with specific simulations. *Pragmatics and Cognition*, 17, 1–42.
- Thomas, J. C., Jr., (1974). An analysis of behavior in the hobbits-orcs problem. *Cognitive Psychology*, 6, 257–269.
- Weisberg, R. W., & Alba, J. W. (1981). An examination of the alleged role of “fixation” in the solution of several “insight” problems. *Journal of Experimental Psychology: General*, 110, 169–192.
- Weiser, M., & Shertz, J. (1983). Programming problem representation in novice and expert programmers. *International Journal of Man-Machine Studies*, 19, 391–398.
- Wertheimer, M. (1959). *Productive thinking* (Rev. ed.). Chicago, IL: University of Chicago Press.
- Winn, W. (1989). The design and use of instructional graphics. In H. Mandl & J. R. Levin (Eds.), *Knowledge acquisition from text and pictures* (pp. 125–144). Amsterdam, Netherlands: Elsevier

On the Distinction Between Rationality and Intelligence: Implications for Understanding Individual Differences in Reasoning

Keith E. Stanovich

Abstract

A concern for individual differences has been missing from the Great Rationality Debate in cognitive science—the debate about how much irrationality to attribute to human cognition. There are individual differences in rational thinking that are less than perfectly correlated with individual differences in intelligence because intelligence and rationality occupy different conceptual locations in models of cognition. A tripartite extension of currently popular dual-process theories is presented in this chapter that illustrates how intelligence and rationality are theoretically separate concepts. The chapter concludes by showing how this tripartite model of mind, taken in the context of studies of individual differences, can help to resolve the Great Rationality Debate.

Key Words: rationality, intelligence, reasoning, individual differences

Introduction

In psychology and among the lay public alike, assessments of intelligence and tests of cognitive ability are taken to be the sine qua non of good thinking. Critics of these instruments often point out that IQ tests fail to assess many domains of psychological functioning that are essential. For example, many largely noncognitive domains such as socioemotional abilities, creativity (Smith & Ward, Chapter 23), empathy, and interpersonal skills are almost entirely unassessed by tests of cognitive ability. However, even these common critiques of intelligence tests often contain the unstated assumption that although intelligence tests miss certain key noncognitive areas, they encompass most of what is important cognitively. In this chapter, I will challenge this tacit assumption by arguing that certain very important classes of individual differences in thinking are ignored if only intelligence-related variance is the focus. Many of these classes of individual differences that are missing from IQ tests are those relating to rational thought (Chater & Oaksford,

Chapter 2). In this chapter, I will illustrate how a comprehensive assessment of individual differences in reasoning skills will necessitate the theoretical and empirical differentiation of the concepts of intelligence and rational thinking.

The discussion in this chapter will begin by showing how differing definitions of rationality frame what is known as the Great Rationality Debate in cognitive science. That debate concerns how much irrationality to attribute to human cognition. I will argue that a concern for individual differences has been missing from this debate because we failed to appreciate that there are individual differences in rational thought as well as intelligence. This is easier to appreciate when we realize that intelligence and rationality occupy somewhat different conceptual locations within most models of cognition. Thus, I present a generic model of the mind that is an extension of currently popular dual-process theories (Evans, Chapter 8), and I situate both intelligence and rationality within this model and show how they dissociate, both conceptually and empirically.

I conclude the chapter by showing how this tripartite model of mind, taking in the context of studies of individual differences, can help to resolve the Great Rationality Debate.

The Concept of Rational Thought in Cognitive Science and Philosophy

The term *rationality* has a strong and a weak sense. The strong sense of the term is the one used in cognitive science, and it will be the one used throughout this chapter. However, a weaker sense of the term has sometimes influenced—and hence confused—arguments in the so-called Great Rationality Debate in cognitive science. The influence of the weak sense of the term has also impeded investigation into individual differences in rational thought.

Dictionary definitions of rationality tend to be of the weak sort—often seeming quite lame and unspecific (“the state or quality of being in accord with reason”). The meaning of rationality in modern cognitive science (the strong sense) is, in contrast, much more specific and prescriptive than this. The weak definitions of rationality derive from a categorical notion of rationality tracing to Aristotle (humans as the only animals who base actions on reason). As de Sousa (2007) has pointed out, such a notion of rationality as “based on reason” has as its opposite not irrationality but *arationality* (outside the domain of reason). Aristotle’s characterization is categorical—the behavior of entities is either based on thought or it is not. Animals are either rational or arational. In this conception, humans are rational, but other animals are not. There is no room for individual differences in rational thinking *among* humans in this view.

In its stronger sense (the sense employed in most of cognitive science and in this chapter) rational thought is a normative notion (Chater & Oaksford, Chapter 2; Griffiths, Tenenbaum, & Kemp, Chapter 3). Its opposite is irrationality, not arationality. Normative models of optimal judgment and decision making define perfect rationality in the noncategorical view employed in cognitive science. Rationality (and irrationality) comes in degrees defined by the distance of the thought or behavior from the optimum defined by a normative model. De Sousa (2007) points out that the notion of rationality in Aristotle’s sense cannot be normative, but in the strong sense of cognitive science, it is. Other animals may be arational, but only humans can be irrational. As de Sousa (2007) puts it, “if human beings can indeed be described as rational animals, it is precisely in virtue of the fact

that humans, of all the animals, are the only ones capable of irrational thoughts and actions” (p. 7).

Hurley and Nudds (2006) make a similar point when they argue that, for a strong sense of the term: “ironically, rationality requires the possibility that the animal might err. It can’t be automatically right, no matter what it does.... when we say that an agent has acted rationally, we imply that it would have been a mistake in some sense to have acted in certain different ways. It can’t be the case that anything the agent might do would count as rational. This is normativity in a quite weak sense” (p. 2). The weak sense they are referring to is an Aristotelian (categorical) sense, and no cognitive scientist is using rationality in this sense when claiming that an experiment has demonstrated human irrationality.

When a cognitive scientist terms a behavior irrational, he or she means that the behavior departs from the optimum prescribed by a particular normative model. The scientist is not implying that no thought or reason was behind the behavior. Some of the hostility that has been engendered by experimental claims of human irrationality no doubt derive from a (perhaps tacit) influence of the Aristotelian view—the thought that cognitive psychologists are saying that certain people are somehow less than human when they are said to behave irrationally. But in using the strong sense of the term *rationality*, most cognitive scientists are saying no such thing.¹

Some of the heat in the Great Rationality Debate is no doubt caused by reactions to the term *irrationality* being applied to humans. As mentioned, lingering associations with the Aristotelian categorical view make charges of irrationality sound more cutting than they actually are when in the context of cognitive science research. When we find a behavioral pattern that is less than optimally rational, we could easily say that it is “less than perfectly rational” rather than that it is irrational—with no loss of meaning. Perhaps if this had been the habit in the literature, the rationality debate in cognitive science would not have become so heated. Such an emphasis also highlights the theme of this chapter—that there are indeed individual differences in rational thought and that understanding the nature of these differences might have important theoretical implications.

Cognitive scientists recognize two types of rationality: epistemic and instrumental. *Epistemic rationality* concerns how well beliefs map onto the actual structure of the world. It is sometimes called theoretical rationality or evidential rationality (see

Audi, 1993, 2001; Foley, 1987; Harman, 1995; Manktelow, 2004; Over, 2004).

The simplest definition of *instrumental rationality* is as follows: behaving in the world so that you get exactly what you most want, given the resources (physical and mental) available to you. Somewhat more technically, we could characterize instrumental rationality as the optimization of the individual's goal fulfillment. Economists and cognitive scientists have refined the notion of optimization of goal fulfillment into the technical notion of expected utility. The model of rational judgment used by decision scientists (Chater & Oaksford, Chapter 2; LeBoeuf & Shafir, Chapter 16) is one in which a person chooses options based on which option has the largest expected utility² (see Baron, 2008; Dawes, 1998; Hastie & Dawes, 2010; Wu, Zhang, & Gonzalez, 2004). One of the fundamental advances in the history of modern decision science was the demonstration that if people's preferences follow certain patterns (the so-called axioms of choice—things like transitivity and freedom from certain kinds of context effects), then they are behaving as if they are maximizing utility; they are acting to get what they most want (Edwards, 1954; Gilboa, 2010; Jeffrey, 1983; Luce & Raiffa, 1957; Savage, 1954; von Neumann & Morgenstern, 1944). This is what makes people's degrees of rationality measurable by the experimental methods of cognitive science. Although it is difficult to assess utility directly, it is much easier to assess whether one of the axioms of rational choice is being violated. This is much like our judgments at a sporting event, where, for example, it might be difficult to discern whether a quarterback has put the ball perfectly on the money, but it is not difficult at all to detect a bad throw.

In fact, in many domains of life this is often the case as well. It is often difficult to specify what the very *best* response might be, but performance *errors* are much easier to spot. Essayist Neil Postman (1988) has argued, for instance, that educators and other advocates of good thinking might adopt a stance more similar to that of physicians or attorneys. He points out that doctors would find it hard to define "perfect health" but, despite this, they are quite good at spotting disease. Likewise, lawyers are much better at spotting injustice and lack of citizenship than defining "perfect justice" or ideal citizenship. Postman argues that, like physicians and attorneys, educators might best focus on instances of poor thinking which are much easier to identify as opposed to trying to define ideal thinking. The

literature on the psychology of rationality has followed this logic in that the empirical literature has focused on identifying thinking errors, just as physicians focus on disease. Degrees of rationality can be assessed in terms of the number and severity of such cognitive biases that individuals display. Conversely, *failure* to display a cognitive bias becomes a measure of rational thought.

The Great Rationality Debate in Cognitive Science

A substantial research literature—one comprising literally hundreds of empirical studies conducted over several decades—has firmly established that people's responses sometimes deviate from the performance considered normative on many reasoning tasks. For example, people assess probabilities incorrectly, they test hypotheses inefficiently, they violate the axioms of utility theory, they do not properly calibrate degrees of belief, their choices are affected by irrelevant context, they ignore the alternative hypothesis when evaluating data, and they display numerous other information processing biases (Baron, 2008; Bazerman & Moore, 2008; Evans, 2007; Gilovich, Griffin, & Kahneman, 2002; Kahneman & Tversky, 2000; Shafir & LeBoeuf, 2002; Stanovich, 2009b, 2011). Demonstrating that descriptive accounts of human behavior diverged from normative models was a main theme of the heuristics and biases research program inaugurated by Kahneman and Tversky in the early 1970s (Kahneman & Tversky, 1972, 1973; Tversky & Kahneman, 1974).

Researchers working in the heuristics and biases tradition tend to be so-called Meliorists (see Bishop & Trout, 2005; Doherty, 2003; Larrick, 2004; Stanovich, 1999, 2004). They assume that human reasoning is not as good as it could be, and that thinking could be improved. The dictionary definition of *meliorism* is "the doctrine that the world tends to become better or may be made better by human effort." Thus, a Meliorist is one who feels that education and the provision of information could help make people more rational—could help them more efficiently further their goals and to bring their beliefs more in line with the actual state of the world.³ Stated this way, Meliorism seems to be an optimistic doctrine, and in one sense it is. But this optimistic part of the Meliorist message derives from the fact that Meliorists see a large gap between normative models of rational responding and descriptive models of what people actually do. Emphasizing the gap, of course, entails that

Meliorists will be attributing a good deal of irrationality to human cognition.

Over the last two decades, an alternative interpretation of the findings from the heuristics and biases research program has been championed. Contributing to this alternative interpretation have been evolutionary psychologists, adaptationist modelers, and ecological theorists (Anderson, 1990; Cosmides & Tooby, 1996; Gigerenzer, 2007; Marewski, Gaissmaier, & Gigerenzer, 2010; Oaksford & Chater, 2007). They have reinterpreted the modal response in most of the classic heuristics and biases experiments as indicating an optimal information processing adaptation on the part of the subjects. It is argued by these investigators that the research in the heuristics and biases tradition has not demonstrated human irrationality at all. This group of theorists—who argue that an assumption of maximal human rationality is the proper default position to take—have been termed the Panglossians (Stanovich, 1999). This position posits no difference between descriptive and normative models of performance because human performance is actually normative.

The contrasting positions of the Panglossians and Meliorists define the differing poles in what has been termed the Great Rationality Debate in cognitive science—the debate about how much irrationality to attribute to human cognition. This debate has generated a very substantial literature of often heated arguments (Cohen, 1981; Doherty, 2003; Edwards & von Winterfeldt, 1986; Evans & Over, 1996, 2010; Gigerenzer, 1996; Jungermann, 1986; Kahneman & Tversky, 1983, 1996; Koehler, 1996; Koehler & James, 2009, 2010; Krueger & Funder, 2004; Kuhberger, 2002; Lee, 2006; Samuels & Stich, 2004; Stanovich, 1999, 2004, 2010; Stanovich & West, 2000; Stein, 1996; Stich, 1990; Vranas, 2000). Tetlock and Mellers (2002) have noted that “the debate over human rationality is a high-stakes controversy that mixes primordial political and psychological prejudices in combustible combinations” (p. 97). The great debate about human rationality is a “high-stakes controversy” because it involves nothing less than the models of human nature that underlie economics, moral philosophy, and the personal theories (folk theories) we use to understand the behavior of other humans. For example, a very influential part of the Panglossian camp is represented by the mainstream of the discipline of economics, which is notable for using strong rationality assumptions as fundamental tools.

Evolution Does Not Guarantee Human Rationality

An Aristotelian view of rationality has no room for individual differences in rational thinking *between* humans. In this view, humans are the unique animals who act based on reason. Thus, all humans are rational—and all are equally so. However, once we move from this view to the normative conception of rationality, we open up room for individual differences. The maximizing notions of rational thought and action in decision science potentially array individuals on a continuum based on the distance of their behavior from the normative model.

We might ask, however, whether—aside from holding an Aristotelian view—there is any other reason to be a Panglossian. One assumption that often draws people to a Panglossian view of human rationality is the thought that evolution would have guaranteed that our cognition is fully rational. This is a mistaken view. There are a number of reasons why evolution would not be expected to guarantee perfect human rationality. One reason is that rationality is defined in terms of maximization (for example, in the case of instrumental rationality, maximizing the expected utility of actions). In contrast to maximization, natural selection works on a “better than” principle. As Dawkins puts it, “Natural selection chooses the better of present available alternatives.... The animal that results is not the most perfect design conceivable, nor is it merely good enough to scrape by. It is the product of a historical sequence of changes, each one of which represented, at best, the better of the alternatives that happened to be around at the time” (p. 46, 1982). In short, the variation and selective retention logic of evolution “designs” (Dennett, 1995) for the reproductive advantage of one organism over the next, not for the optimality of any one characteristic (including rationality). It has been said that evolution should be described as the survival of the *fitter* rather than as the survival of the fittest.

Evolution proceeds to increase the reproductive fitness of genes, not to increase the rationality of humans (Stanovich, 2004; Stanovich & West, 2003). Increases in fitness do not always entail increases in rationality. Take, for example, the domain of beliefs. Beliefs need not always track the world with maximum accuracy in order for fitness to increase (Stich, 1990). Thus, evolution does not guarantee perfect epistemic rationality. For example, evolution might fail to select out epistemic mechanisms of high accuracy when they are costly

in terms of organismic resources (for example, in terms of memory, energy, or attention). An additional reason that belief-forming mechanisms might not be maximally truth preserving is that:

a very cautious, risk-aversive inferential strategy—one that leaps to the conclusion that danger is present on very slight evidence—will typically lead to false beliefs more often, and true ones less often, than a less hair-trigger one that waits for more evidence before rendering a judgment. Nonetheless, the unreliable, error-prone, risk-aversive strategy may well be favored by natural selection. For natural selection does not care about truth; it cares only about reproductive success. (p. 62, Stich, 1990)

It is likewise in the domain of goals and desires. As has become clear from recent research on the topic of affective forecasting, people are remarkably bad at making choices that make themselves happy (Gilbert, 2006; Kahneman et al., 2006; Wilson & Gilbert, 2005). This should be no surprise. The reason we have pleasure circuits in our brains is to encourage us to do things (survive and reproduce, help kin) that propagate our genes. The pleasure centers were not designed to maximize the amount of time we are happy.

The instrumental rationality of humans is not guaranteed by evolution for two further reasons. First, many genetic goals that have been lodged in our brain no longer serve our ends because the environment has changed (Richerson & Boyd, 2005). For example, thousands of years ago, humans needed as much fat as they could get in order to survive. More fat meant longer survival and because few humans survived beyond their reproductive years, longevity translated directly into more opportunities for gene replication. In short, our mechanisms for storing and utilizing energy evolved in times when fat preservation was efficacious. These mechanisms no longer serve the goals of people in our modern technological society where there is a McDonald's on practically every corner—the goals underlying these mechanisms have become detached from their evolutionary context.

Finally, rational standards for assessing human behavior are social and cultural products that are preserved and stored independently of the genes. The development of probability theory, concepts of empiricism, logic, and scientific thinking throughout the centuries have provided humans with conceptual tools to aid in the formation and revision of belief and in their reasoning about action (Chater & Oaksford, Chapter 2; Griffiths et al.,

Chapter 3; Dunbar & Klahr, Chapter 35). They represent the cultural achievements that foster greater human rationality (Thagard & Nisbett, 1983). As societies evolve, they produce more of the cultural tools of rationality and these tools become more widespread in the population. Thus, the cultural evolution of rational standards (Thagard & Nisbett, 1983) is apt to occur at a pace markedly faster than that of human evolution—providing ample opportunity for mental strategies of utility maximization to dissociate from local genetic fitness maximization. In summary, a consideration of our evolutionary history should not lead one to a Panglossian view of human rationality.

A reconciliation of the views of the Panglossians and Meliorists is possible, however, if we take three scientific steps. First, we must consider data patterns long ignored in the heuristics and biases literature—individual differences on rational thinking tasks. Second, we must understand the empirical patterns obtained through the lens of a modified dual-process theory (Evans, Chapter 8) and of evolutionary theory. Thirdly, we must distinguish the concepts of rationality and intelligence in cognitive theory. Subsequent sections of this chapter develop each of these points.

Individual Differences in the Great Rationality Debate

Dozens of empirical studies have shown that there are few tasks in the heuristics and biases literature where all untutored laypersons give the same response. What has largely been ignored is that—although the average person in the classic heuristics and biases experiments might well display an overconfidence effect, underutilize base rates, ignore P(D/-H), violate the axioms of utility theory, choose P and Q in the selection task, commit the conjunction fallacy, and so on—on each of these tasks, *some people give the standard normative response* (Bruine de Bruin, Parker, & Fischhoff, 2007; Cokely & Kelley, 2009; Del Missier, Mantyla, & Bruine de Bruin, 2010; Dohmen et al., 2009; Finucane & Gullion, 2010; Frederick, 2005; Klaczynski, 2001; Oechssler, Roider, & Schmitz, 2009; Stanovich & West, 1998b, 1998c, 1999, 2000, 2008b; West et al., 2008). What has been ignored in the Great Rationality Debate is individual differences. For example, in knowledge calibration studies, although the mean performance level of the entire sample may be represented by a calibration curve that indicates overconfidence, almost always some

people do display near perfect calibration. Likewise, in probabilistic assessment, while the majority of subjects might well ignore the noncausal base-rate evidence, a minority of subjects often makes use of this information in exactly the way prescribed by Bayes' theorem. A few people even respond correctly on the notoriously difficult abstract selection task (Evans, Newstead, & Byrne, 1993; Stanovich & West, 1998a, 2008b).

In short, some people give the response traditionally considered normative, and others do not. There is variability in responding on all of these tasks. So when Panglossians and heuristics and biases researchers argue about the normative appropriateness of a particular response, whoever eventually prevails in the dispute—both sides have been guilty of glossing over individual differences. In short, it is incorrect to imply that people uniformly display a particular rational or irrational response pattern. A particular experiment might instead be said to show that the average person, or perhaps the modal person, displays optimal or suboptimal thinking. Other people, often a minority to be sure, display the opposite style of thinking.

In light of these empirical data, it is puzzling that Panglossians would presumably accept the existence of individual differences in intelligence but not rationality. This is possible, however, if intelligence and rationality are two different things conceptually. In the remainder of this chapter, I will show that the Panglossians are correct in one of their assumptions but incorrect in another. Conceptually, intelligence and rational thinking are indeed two different things. But—contra Panglossian assumptions—the latter as well as the former displays substantial individual differences.

Discussions of intelligence often go off the rails at the very beginning by failing to set the concept within a general context of cognitive functioning—thus inviting the default assumption that intelligence is the central feature of the mind. I will try to preclude this natural default by outlining a model of the mind and then placing intelligence within it. The generic models of the mind developed by cognitive scientists often give short shrift to a question that the public is intensely interested in: How and why do people *differ* from each other in their thinking? In an attempt to answer that question, I am going to present a gross model of the mind that is true to modern cognitive science but that emphasizes individual differences in ways that are somewhat new. The model builds on a current consensus

view of cognition termed dual-process theory (see Evans, Chapter 8, for a more detailed discussion).

From Dual-Process Theory to a Tripartite Model of Mind

The idea that the brain is composed of many different subsystems (see Aunger & Curtis, 2008) has recurred in conceptualizations in many different disciplines—from the society of minds view in artificial intelligence (Minsky, 1985); to Freudian analogies (Ainslie, 1982); to discussions of the concept of multiple selves in philosophy, economics, and decision science (Ainslie, 2001; Schelling, 1984). In fact, the notion of many different systems in the brain is by no means new. Plato (1945) argued that “we may call that part of the soul whereby it reflects, rational; and the other, with which it feels hunger and thirst and is distracted by sexual passion and all the other desires, we will call irrational appetite, associated with pleasure in the replenishment of certain wants” (p. 137).

What is new, however, is that cognitive scientists are beginning to understand the biology and cognitive structure of these systems (Evans & Frankish, 2009; see Morrison & Knowlton, Chapter 6; Green & Dunbar, Chapter 7) and are beginning to posit some testable speculations about their evolutionary and experiential origins. I will build on the current consensus that the functioning of the brain can be characterized by two different types of cognition having somewhat different functions and different strengths and weaknesses. There is a wide variety of evidence that has converged on the conclusion that some type of dual-process notion is needed in a diverse set of specialty areas not limited to cognitive psychology, social psychology, naturalistic philosophy, decision theory, and clinical psychology (Evans, 2003, 2008, 2010; Frankish, 2004; Lieberman, 2007, 2009; Schneider & Chein, 2003; Sloman, 1996, 2002; Smith & Decoster, 2000; Stanovich, 1999). Evolutionary theorizing and neurophysiological work also have supported a dual-process conception (Camerer, Loewenstein, & Prelec, 2005; Frank, Cohen, & Sanfey, 2009; Lieberman, 2009; McClure, Laibson, Loewenstein, & Cohen, 2004; Prado & Noveck, 2007; Reber, 1993; Toates, 2005, 2006). In fact, a dual-process view was implicit within the early writings in the groundbreaking heuristics and biases research program. As Kahneman (2000) notes, “Tversky and I always thought of the heuristics and biases approach as a two-process theory” (p. 682).

Just how ubiquitous are dual-process models in psychology and related fields is illustrated in

Table 22.1, which lists a variety of such theories that have appeared during the last couple of decades. Some common terms for the dual processes are listed in Table 22.1. My purpose here is not to adjudicate the differences among these models. Instead, I will gloss over differences and instead start with

a model that emphasizes the family resemblances. Evans (Chapter 8) provides a much more nuanced discussion.

The family resemblances extend to the names for the two classes of process. The terms *heuristic* and *analytic* are two of the oldest and most popular (see

Table 22.1 Some Alternative Terms for Type 1 and Type 2 Processing Used by Various Theorists

Theorist	Type 1	Type 2
Bargh & Chartrand (1999)	Automatic processing	Conscious processing
Bazerman, Tenbrunsel, & Wade-Benzoni, (1998)	Want self	Should self
Bickerton (1995)	Online thinking	Offline thinking
Brainerd & Reyna (2001)	Gist processing	Analytic processing
Chaiken et al. (1989)	Heuristic processing	Systematic processing
Evans (1984, 1989)	Heuristic processing	Analytic processing
Evans & Over (1996)	Tacit thought processes	Explicit thought processes
Evans & Wason (1976); Wason & Evans (1975)	Type 1 processes	Type 2 processes
Fodor (1983)	Modular processes	Central processes
Gawronski & Bodenhausen (2006)	Associative processes	Propositional processes
Haidt (2001)	Intuitive system	Reasoning system
Johnson-Laird (1983)	Implicit inferences	Explicit inferences
Kahneman & Frederick (2002, 2005)	Intuition	Reasoning
Lieberman (2003)	Reflexive system	Reflective system
Loewenstein (1996)	Visceral factors	Tastes
Metcalfe & Mischel (1999)	Hot system	Cool system
Norman & Shallice (1986)	Contention scheduling	Supervisory attentional system
Pollock (1991)	Quick and inflexible modules	Intellection
Posner & Snyder (1975)	Automatic activation	Conscious processing
Reber (1993)	Implicit cognition	Explicit learning
Shiffrin & Schneider (1977)	Automatic processing	Controlled processing
Sloman (1996)	Associative system	Rule-based system
Smith & DeCoster (2000)	Associative processing	Rule-based processing
Strack & Deutsch (2004)	Impulsive system	Reflective system
Thaler & Shefrin (1981)	Doer	Planner
Toates (2006)	Stimulus-bound	Higher order
Wilson (2002)	Adaptive unconscious	Conscious

Evans, 1984, 1989). However, to attenuate the proliferation of nearly identical theories, I suggested the more generic terms System 1 and System 2 in a previous book (Stanovich, 1999). Although these terms have become popular, there is an infelicitousness to the System 1/System 2 terminology. Such terminology seems to connote that the two processes in dual-process theory map explicitly to two distinct brain systems. This is a stronger assumption than most theorists wish to make. Additionally, both Evans (2008, 2009, 2010; Chapter 8, this volume) and Stanovich (2004, 2011) have discussed how terms such as *System 1* or *heuristic system* are really misnomers because they imply that what is being referred to is a singular system. In actuality, the term used should be plural because it refers to a *set* of systems in the brain that operate autonomously in response to their own triggering stimuli and are not under higher level cognitive control. I have suggested (Stanovich, 2004) the acronym TASS (standing for The Autonomous Set of Systems) to describe what is in actuality a heterogeneous set.

Using the acronym TASS was a step forward in clearing up some of the confusion surrounding autonomous processes. For similar reasons, Evans (2008, 2009, Chapter 8; see also Samuels, 2009) has suggested a terminology of Type 1 processing versus Type 2 processing. The Type 1/Type 2 terminology captures better than previous terminology that a dual-*process* theory is not necessarily a dual-*system* theory (see Evans, 2008, 2009, for an extensive discussion). For these reasons, I will rely most heavily on the Type 1/Type 2 terminology. An even earlier terminology due to Evans (1984, 1989)—heuristic versus analytic processing—will also be employed on occasions when it is felicitous.

The defining feature of Type 1 processing is its autonomy—the execution of Type 1 processes is mandatory when their triggering stimuli are encountered, and they are not dependent on input from high-level control systems. Autonomous processes have other correlated features—their execution is rapid, they do not put a heavy load on central processing capacity, they tend to operate in parallel without interfering with themselves or with Type 2 processing—but these other correlated features are not defining. Autonomous processes would include behavioral regulation by the emotions; the encapsulated modules for solving specific adaptive problems that have been posited by evolutionary psychologists; processes of implicit learning; and the automatic firing of overlearned associations. Type 1

processes conjoin the properties of automaticity, quasi-modularity, and heuristic processing as these constructs have been variously discussed in cognitive science (Barrett & Kurzban, 2006; Carruthers, 2006; Coltheart, 1999; Evans, 2008, 2009; Moors & De Houwer, 2006; Samuels, 2005, 2009; Shiffrin & Schneider, 1977; Sperber, 1994).

It is important to emphasize that Type 1 processing is not limited to modular subprocesses that meet all of the classic Fodorian (1983) criteria. Type 1 processing encompasses processes of unconscious implicit learning and conditioning. Also, many rules, stimulus discriminations, and decision-making principles that have been practiced to automaticity (e.g., Kahneman & Klein, 2009; Shiffrin & Schneider, 1977) are processed in a Type 1 manner. This learned information can sometimes be just as much a threat to rational behavior as are evolutionary modules that fire inappropriately in a modern environment. Rules learned to automaticity can be overgeneralized—they can autonomously trigger behavior when the situation is an exception to the class of events they are meant to cover (Arkes & Ayton, 1999; Hsee & Hastie, 2006).

Type 2 processing is nonautonomous. Type 2 processing contrasts with Type 1 processing on each of the correlated properties that define the latter. It is relatively slow and computationally expensive. Many Type 1 processes can operate at once in parallel, but Type 2 processing is largely serial. Type 2 processing is often language based, but it is not necessarily so. One of the most critical functions of Type 2 processing is to override Type 1 processing. This is sometimes necessary because autonomous processing has heuristic qualities. It is designed to get the response into the right ballpark when solving a problem or making a decision, but it is not designed for the type of fine-grained analysis called for in situations of unusual importance (financial decisions, fairness judgments, employment decisions, legal judgments, etc.). Heuristics depend on benign environments. In hostile environments, they can be costly (see Hilton, 2003; Over, 2000; Stanovich, 2004, 2009b). A benign environment means one that contains useful (that is, diagnostic) cues that can be exploited by various heuristics (for example, affect-triggering cues, vivid and salient stimulus components, convenient and accurate anchors). Additionally, for an environment to be classified as benign, it also must contain no other individuals who will adjust their behavior to exploit those relying only on heuristics. In contrast,

a hostile environment for heuristics is one in which there are few cues that are usable by heuristic processes or there are misleading cues (Kahneman & Klein, 2009). Another way that an environment can turn hostile for a heuristic processor is if other agents discern the simple cues that are being used and the other agents start to arrange the cues for their own advantage (for example, advertisements, or the deliberate design of supermarket floor space to maximize revenue).

All of the different kinds of Type 1 processing (processes of emotional regulation, Darwinian modules, associative and implicit learning processes) can produce responses that are irrational in a particular context if not overridden. For example, often humans act as cognitive misers (see Stanovich, 2009b) by engaging in attribute substitution (Kahneman & Frederick, 2002)—the substitution of an easy-to-evaluate characteristic for a harder one, even if the easier one is less accurate. For example, the cognitive miser will substitute the less effortful attributes of vividness or affect for the more effortful retrieval of relevant facts (Kahneman, 2003; Li & Chapman, 2009; Slovic & Peters, 2006; Wang, 2009). But when we are evaluating important risks—such as the risk of certain activities and environments for our children—we do not want to substitute vividness for careful thought about the situation. In such situations, we want to employ Type 2 override processing to block the attribute substitution of the cognitive miser.

To override Type 1 processing, Type 2 processing must display at least two related capabilities. One is the capability of interrupting Type 1 processing and suppressing its response tendencies. Type 2 processing thus involves inhibitory mechanisms of the type that have been the focus of work on executive functioning (Aron, 2008; Best, Miller, & Jones, 2009; Hasher, Lustig, & Zacks, 2007; Miyake et al., 2000; Zelazo, 2004). But the ability to suppress Type 1 processing gets the job only half done. Suppressing one response is not helpful unless there is a better response available to substitute for it. Where do these better responses come from? One answer is that they come from processes of hypothetical reasoning and cognitive simulation that are a unique aspect of Type 2 processing (Johnson-Laird, Chapter 9). When we reason hypothetically, we create temporary models of the world and test out actions (or alternative causes) in that simulated world. To reason hypothetically we must, however, have one critical cognitive capability—we must be

able to prevent our representations of the real world from becoming confused with representations of imaginary situations. The so-called cognitive decoupling operations are the central feature of Type 2 processing that make this possible, and they have implications for how we conceptualize both intelligence and rationality.

In a much-cited article, Leslie (1987) modeled pretense by positing a so-called secondary representation (see Perner, 1991) that was a copy of the primary representation but that was decoupled from the world so that it could be manipulated—that is, be a mechanism for simulation. The important issue for our purposes is that decoupling secondary representations from the world and then maintaining the decoupling while simulation is carried out is a Type 2 processing operation. It is computationally taxing and greatly restricts the ability to conduct any other Type 2 operation simultaneously. In fact, decoupling operations might well be a major contributor to a distinctive Type 2 property: its seriality.

Figure 22.1 represents a preliminary model of mind, based on what has been outlined thus far, with one important addition. The addition stems from the fact that instructions to initiate override of Type 1 processing (and to initiate simulation activities) must be controlled by cognitive machinery at a higher level than the decoupling machinery itself. Type 2 processing needs to be understood in terms of two levels of cognitive control—what are termed in Figure 22.1 the algorithmic level and the reflective level. There I have presented the tripartite proposal in the spirit of Dan Dennett's (1996) book *Kinds of Minds* where he used that title to suggest that within the brain of humans are control systems of very different types—different kinds of minds. I have labeled the traditional source of Type 1 processing as the autonomous mind but differentiated Type 2 processing into the algorithmic mind and the reflective mind. The autonomous mind can be overridden by algorithmic-level mechanisms; but override itself is initiated by higher level control. That is, the algorithmic level is conceptualized as subordinate to the higher level goal states and epistemic thinking dispositions of the reflective mind.

Individual Differences Within the Tripartite Model of Mind

Psychometricians have long distinguished typical performance situations from optimal (sometimes termed *maximal*) performance situations (see Ackerman, 1994, 1996; Ackerman & Heggestad,

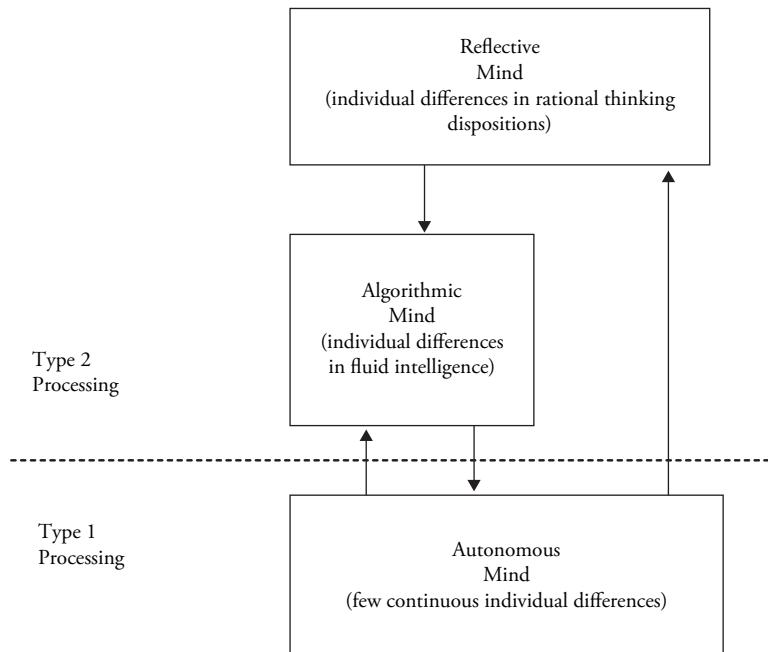


Fig. 22.1 The tripartite structure and the locus of individual differences.

1997; Ackerman & Kanfer, 2004; see also Cronbach, 1949; Matthews, Zeidner, & Roberts, 2002; Sternberg, Grigorenko, & Zhang, 2008). Typical performance situations are unconstrained in that no overt instructions to maximize performance are given, and the task interpretation is determined to some extent by the participant. The goals to be pursued in the task are left somewhat open. The issue is what a person would typically do in such a situation, given few constraints. Typical performance measures are measures of the reflective mind—they assess in part goal prioritization and epistemic regulation. In contrast, optimal performance situations are those where the task interpretation is determined externally. The person performing the task is instructed to maximize performance. Thus, optimal performance measures examine questions of the efficiency of goal pursuit—they capture the processing efficiency of the algorithmic mind. All tests of intelligence or cognitive aptitude are optimal performance assessments, whereas measures of critical or rational thinking are often assessed under typical performance conditions.

The difference between the algorithmic mind and the reflective mind is captured in another well-established distinction in the measurement of individual differences: the distinction between cognitive ability and thinking dispositions. The former are,

as just mentioned, measures of the efficiency of the algorithmic mind. The latter travel under a variety of names in psychology—*thinking dispositions* or *cognitive styles* being the two most popular. Many thinking dispositions concern beliefs, belief structure and, importantly, attitudes toward forming and changing beliefs. Other thinking dispositions that have been identified concern a person's goals and goal hierarchy. Examples of some thinking dispositions that have been investigated by psychologists are actively open-minded thinking, need for cognition (the tendency to think a lot), consideration of future consequences, need for closure, superstitious thinking, and dogmatism (Cacioppo et al., 1996; Kruglanski & Webster, 1996; Norris & Ennis, 1989; Schommer-Aikins, 2004; Stanovich, 1999, 2009b; Sternberg, 2003; Sternberg & Grigorenko, 1997; Strathman et al., 1994).

The types of cognitive propensities that these thinking disposition measures reflect are the tendency to collect information before making up one's mind, the tendency to seek various points of view before coming to a conclusion, the disposition to think extensively about a problem before responding, the tendency to calibrate the degree of strength of one's opinion to the degree of evidence available, the tendency to think about future consequences before taking action, the tendency to explicitly weigh pluses

and minuses of situations before making a decision, and the tendency to seek nuance and avoid absolutism. In short, individual differences in thinking dispositions are assessing variation in people's goal management, epistemic values, and epistemic self-regulation—differences in the operation of reflective mind. They are psychological characteristics that underpin rational thought and action.

The cognitive abilities assessed on intelligence tests are not of this type. They are not about high-level personal goals and their regulation, or about the tendency to change beliefs in the face of contrary evidence, or about how knowledge acquisition is internally regulated when not externally directed. People have indeed come up with *definitions* of intelligence that encompass such things. Theorists often define intelligence in ways that encompass rational action and belief but, nevertheless, *the actual measures of intelligence in use assess only algorithmic-level cognitive capacity*. No current intelligence test that is even moderately used in practice assesses rational thought or behavior (Stanovich, 2002, 2009b).

Figure 22.1 represents the classification of individual differences in the tripartite view. The broken horizontal line represents the location of the key distinction in older, dual-process views. Whereas the reflective and algorithmic minds are characterized by continuous individual differences and substantial variability, there are fewer continuous individual differences in the autonomous mind and less variability (see Kaufman et al., 2010, for a different view). Disruptions to the autonomous mind often reflect damage to cognitive modules that result in very discontinuous cognitive dysfunction such as autism or the agnosias and alexias (Anderson, 2005; Bermudez, 2001; Murphy & Stich, 2000).

Figure 22.1 identifies variation in fluid intelligence (Gf) with individual differences in the efficiency of processing of the algorithmic mind. Fluid intelligence is one component in the Cattell/Horn/Carroll (CHC) theory of intelligence (Carroll, 1993; Cattell, 1963, 1998; Horn & Cattell, 1967). Sometimes called the theory of fluid and crystallized intelligence (symbolized Gf/Gc theory), this theory posits that tests of mental ability tap, in addition to a general factor, a small number of broad factors, of which two are dominant (Geary, 2005; Horn & Noll, 1997; Taub & McGrew, 2004). Fluid intelligence (Gf) reflects reasoning abilities operating across a variety of domains—in particular, novel ones. It is measured by tasks of abstract reasoning

such as figural analogies, Raven matrices, and series completion. Crystallized intelligence (Gc) reflects declarative knowledge acquired from acculturated learning experiences. It is measured by vocabulary tasks, verbal comprehension, and general knowledge measures. Ackerman (1996) discusses how the two dominant factors in the CHC theory reflect a long history of considering two aspects of intelligence: intelligence-as-process (Gf) and intelligence-as-knowledge (Gc).

I have argued that individual differences in fluid intelligence are a key indicator of the variability across individuals in the ability to sustain decoupling operations (Stanovich, 2001, 2009b). Increasingly it is becoming apparent that one of the critical mental operations being tapped by measures of fluid intelligence is the cognitive decoupling operation I have discussed in this chapter. This is becoming clear from converging work on executive function and working memory. Most measures of executive function and working memory are direct or indirect indicators of a person's ability to sustain decoupling operations (Duncan et al., 2008; Engle, 2002; Gray, Chabris, & Braver, 2003; Hasher, Lustig, & Zacks, 2007; Kane, 2003; Lepine, Barrouillet, & Camos, 2005; Salthouse, Atkinson, & Berish, 2003; Salthouse & Pink, 2008; Stanovich, 2011).

Figure 22.1 highlights an important sense in which rationality is a more encompassing construct than intelligence. As previously discussed, to be rational, a person must have well-calibrated beliefs and must act appropriately on those beliefs to achieve goals—both properties of the reflective mind. The person must, of course, have the algorithmic-level machinery that enables him or her to carry out the actions and to process the environment in a way that enables the correct beliefs to be fixed and the correct actions to be taken. Thus, individual differences in rational thought and action can arise because of individual differences in fluid intelligence (the algorithmic mind) or because of individual differences in thinking dispositions (the reflective mind).

The conceptualization in Figure 22.1 has several advantages. First, it conceptualizes intelligence in terms of what intelligence tests actually measure. IQ tests do not attempt to measure directly an aspect of epistemic or instrumental rationality, nor do they examine any thinking dispositions that relate to rationality. It is also clear from Figure 22.1 why rationality and intelligence can become dissociated. Rational thinking depends on thinking dispositions

as well as algorithmic efficiency. Thus, as long as variation in thinking dispositions is not perfectly correlated with fluid intelligence, there is the statistical possibility of dissociations between rationality and intelligence.

In fact, substantial empirical evidence indicates that individual differences in thinking dispositions and intelligence are far from perfectly correlated. Many different studies involving thousands of subjects (e.g., Ackerman & Heggestad, 1997; Austin & Deary, 2002; Baron, 1982; Bates & Shieles, 2003; Cacioppo et al., 1996; Eysenck, 1994; Goff & Ackerman, 1992; Kanazawa, 2004; Kokis et al., 2002; Zeidner & Matthews, 2000) have indicated that measures of intelligence display only moderate to weak correlations (usually less than .30) with some thinking dispositions (e.g., actively open-minded thinking, need for cognition) and near-zero correlations with others (e.g., conscientiousness, curiosity, diligence). Other important evidence supports the conceptual distinction made here between algorithmic cognitive capacity and thinking dispositions. For example, across a variety of tasks from the heuristics and biases literature, it has consistently been found that rational thinking dispositions will predict variance after the effects of general intelligence have been controlled.⁴

The functions of the different levels of control are illustrated more completely in Figure 22.2. There, it is clear that the override capacity itself is a property of the algorithmic mind and it is indicated by the arrow labeled A. However, previous dual-process theories have tended to ignore the higher level cognitive function that initiates the override function in the first place. This is a dispositional property of the reflective mind that is related to rationality. In the model in Figure 22.2, it corresponds to arrow B, which represents (in machine intelligence terms) the call to the algorithmic mind to override the Type 1 response by taking it offline. This is a different mental function than the override function itself (arrow A), and there, the evidence cited earlier indicates that the two functions are indexed by different types of individual differences.

Figure 22.2 represents another aspect of cognition somewhat neglected by previous dual-process theories. Specifically, the override function has loomed large in dual-process theory but less so the simulation process that computes the alternative response that makes the override worthwhile. Figure 22.2 explicitly represents the simulation function as well as the fact that the call to initiate simulation originates in the reflective mind. The decoupling

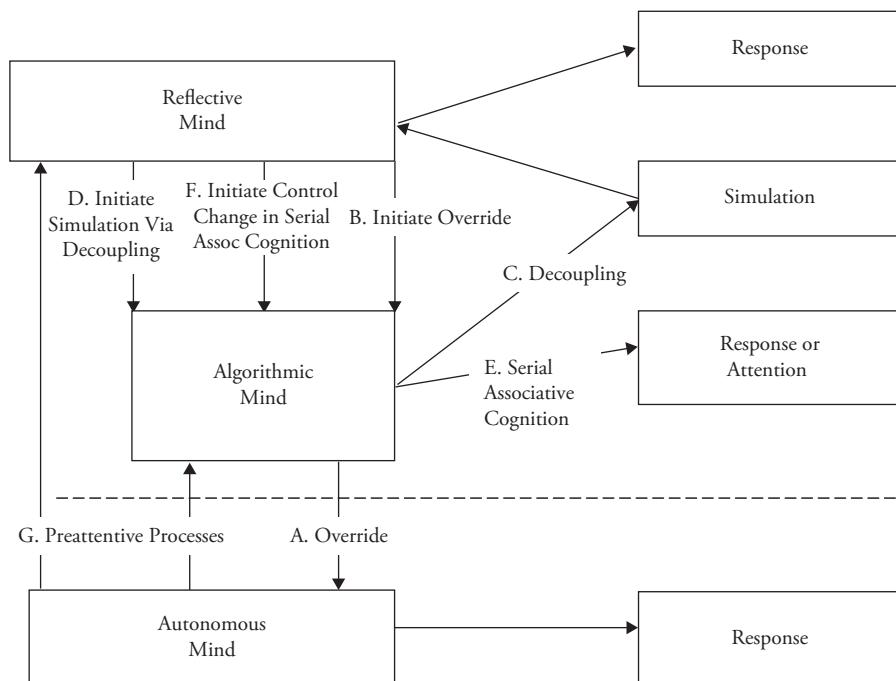


Fig. 22.2 A more complete model of the tripartite structure.

operation (indicated by arrow C) itself is carried out by the algorithmic mind and the call to initiate simulation (indicated by arrow D) by the reflective mind. Again, two different types of individual differences are associated with the initiation call and the decoupling operator—specifically, rational thinking dispositions with the former and fluid intelligence with the latter. Also represented is the fact that the higher levels of control receive inputs from the computations of the autonomous mind (arrow G) via so-called preattentive processes (Evans, 2006, 2007, 2008, 2009). The arrows labeled E and F reflect the decoupling and higher level control of a kind of Type 2 processing (serial associative cognition) that does not involve fully explicit cognitive simulation (see Stanovich, 2011).

Mindware in the Tripartite Model

Knowledge bases, both innate and derived from experience, also importantly bear on rationality. The term *mindware* was coined by Perkins (1995) to refer to the rules, knowledge, procedures, and strategies that a person can retrieve from memory in order to aid decision making and problem solving. Each of the levels in the tripartite model of mind has to access knowledge to carry out its operations,

as illustrated in Figure 22.3. As the Figure indicates, the reflective mind not only accesses general knowledge structures (G_c) but, importantly, accesses the person's opinions, beliefs, and reflectively acquired goal structure. The algorithmic mind accesses microstrategies for cognitive operations and production system rules for sequencing behaviors and thoughts. Finally, the autonomous mind accesses not only knowledge bases that are evolutionary adaptations, but it also retrieves information that has become tightly compiled and available to the autonomous mind due to overlearning and practice.

It is important to note that what is displayed in Figure 22.3 are the knowledge bases that are *unique* to each mind. Algorithmic- and reflective-level processes also receive inputs from the computations of the autonomous mind (see arrow G in Figure 22.2). The mindware available for retrieval, particularly that available to the reflective mind, is in part the product of past learning experiences. The knowledge structures available for retrieval by the reflective mind represent G_c , crystallized intelligence. Recall that G_f , fluid intelligence (intelligence-as-process), is already represented in the Figure 22.2. It is the general computational power of the algorithmic

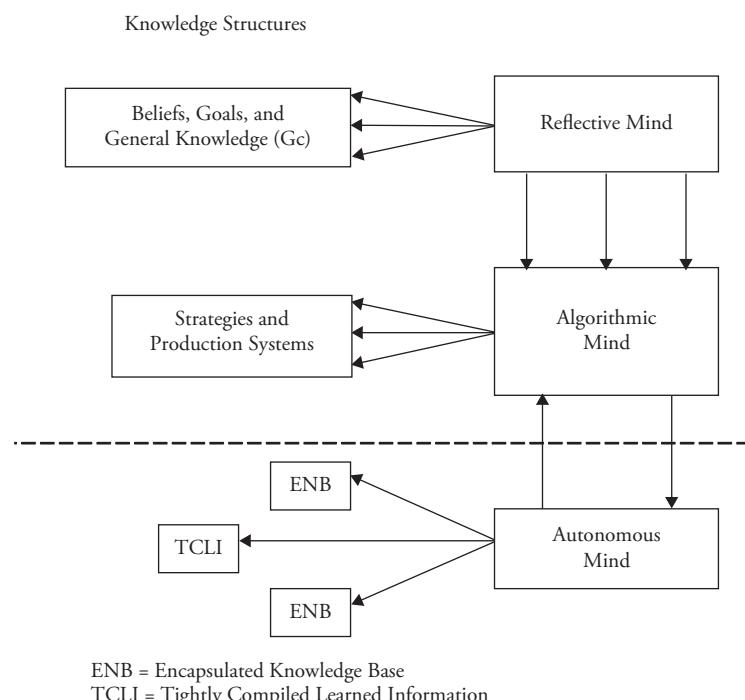


Fig. 22.3 Knowledge structures in the tripartite framework.

mind—importantly exemplified by the ability to sustain cognitive decoupling.

It is important to see how both of the major components of Gf/Gc theory miss critical aspects of rational thought. Fluid intelligence will, of course, have some relation to rationality because it indexes the computational power of the algorithmic mind to sustain decoupling. Because override and simulation are important operations for rational thought, Gf will definitely facilitate rational action in some situations. Nevertheless, the tendency to initiate override (arrow B in Fig. 22.2) and to initiate simulation activities (arrow D in Fig. 22.2) are both aspects of the reflective mind unassessed by intelligence tests, so the tests will miss these components of rationality.

The situation with respect to Gc is a little different. It is true that much of the mindware of rational thought would be classified as crystallized intelligence in the abstract. But is it the kind of crystallized knowledge that is specifically assessed on the tests? The answer is no. The mindware of rational thought is somewhat specialized mindware (see Stanovich, 2009b). It clusters in the domains of probabilistic reasoning (see Griffiths et al., Chapter 3), causal reasoning (see Cheng & Buehner, Chapter 12), and scientific reasoning (see Dunbar & Klahr, Chapter 35). In contrast, the crystallized knowledge assessed on IQ tests is deliberately designed to be nonspecialized. The designers of the tests, in order to make sure the sampling of Gc is fair and unbiased, explicitly attempt to *broadly* sample vocabulary, verbal comprehension domains, and general knowledge. The broad sampling ensures unbiasedness in the test, but it inevitably means that the specific knowledge bases critical to rationality will go unassessed. In short, Gc, as traditionally measured, does not assess individual differences in rationality, and Gf will do so only indirectly and to a mild extent. In short, as measures of rational thinking, IQ tests are radically incomplete.

The Requirements of Rational Thinking and Their Relation to Intelligence

Within the tripartite framework, rationality requires mental characteristics of three different types. Problems in rational thinking arise when cognitive capacity is insufficient to sustain autonomous system override, when the necessity of override is not recognized, or when simulation processes do not have access to the mindware necessary for the synthesis of a better response. The source of these

problems, and their relation to intelligence, help to explain one data trend that has been uncovered—that some rational thinking problems show surprising degrees of dissociation from cognitive ability (Stanovich, 2009b, 2011; Stanovich & West, 2007, 2008a, 2008b; West, Toplak, & Stanovich, 2008). Myside bias, for example, is virtually independent of intelligence (Macpherson & Stanovich, 2007; Sá, Kelley, Ho, & Stanovich, 2005; Stanovich & West, 2007, 2008a, 2008b; Toplak & Stanovich, 2003). Individuals with higher IQs in a university sample are no less likely to process information from an egocentric perspective than are individuals with relatively lower IQs.

Irrational behavior can occur because the right mindware (cognitive rules, strategies, knowledge, and belief systems) is not available to use in decision making. We would expect to see a correlation with intelligence here because mindware gaps most often arise from lack of education or experience. Nevertheless, while it is true that more intelligent individuals learn more things than less intelligent individuals, much knowledge (and many thinking dispositions) relevant to rationality are picked up rather late in life. Explicit teaching of this mindware is not uniform in the school curriculum at any level. That such principles are taught very inconsistently means that some intelligent people may fail to learn these important aspects of critical thinking. In university samples, correlations with cognitive ability have been found to be roughly (in absolute magnitude) in the range of .20–.35 for probabilistic reasoning tasks and scientific reasoning tasks measuring a variety of rational principles (Bruine de Bruin, Parker, & Fischhoff, 2007; Kokis et al., 2002; Parker & Fischhoff, 2005; Sá, West, & Stanovich, 1999; Stanovich & West, 1998b, 1998c, 1998d, 1999, 2000, 2008b; Toplak & Stanovich, 2002). This is again a magnitude of correlation that allows for substantial discrepancies between intelligence and rationality. Intelligence is thus no inoculation against many of the sources of irrational thought. None of these sources are directly assessed on intelligence tests, and the processes that *are* tapped by IQ tests are not highly overlapping with the processes and knowledge that explain variation in rational thinking ability.

Conclusions and Future Directions

The many studies of individual differences on heuristics and biases tasks falsify the most explicit versions of the Panglossian view. People are not all

identically rational. There are individual differences on all of these tasks and the individual differences are not the result of random performance errors (see Stanovich, 1999; Stein, 1996). Instead, they are systematic. If there are such systematic individual differences, it means that at least some people, some of the time, are irrational. Extreme Panglossianism cannot be sustained. However, there is another position in the debate that, like Panglossianism, serves to minimize the attribution of irrationality. Termed the Apologist position (Stanovich, 1999), this view takes very seriously the idea that humans have computational limitations that keep them from being fully rational.

Like the Meliorist, the Apologist accepts the empirical reality and nonspuriousness of normative/descriptive gaps, but the Apologist is much more hesitant to term them instances of irrationality. This is because the Apologist takes the position that to characterize a suboptimal behavior as irrational, it must be the case that the normative model is computable by the individual. If there are computational limitations affecting task performance, then the normative model may not be prescriptive, at least for individuals of low algorithmic capacity. Prescriptive models are usually viewed as specifying how processes of belief formation and decision making should be carried out, given the limitations of the human cognitive apparatus and the situational constraints (e.g., time pressure) with which the decision maker must deal (Baron, 2008). From the Apologist's perspective, the descriptive model is quite close to the prescriptive model and the descriptive/normative gap is attributed to a computational limitation. Although the Apologist admits that performance is suboptimal from the standpoint of the normative model, it is not irrational because there is no prescriptive/descriptive gap.

However, as demonstrated in some of the data patterns I have described in this chapter, the Apologist stratagem will not work for all of the irrational tendencies that have been uncovered in the heuristics and biases literature. This is because many biases are not very strongly correlated with measures of intelligence (algorithmic capacity). Additionally, there is reliable variance in rational thinking found even after cognitive ability is controlled, and that reliable variance is associated with thinking dispositions in theoretically predictable ways. These thinking dispositions reflect control features of the reflective mind that can lead to responses that are more or less rational. They are one of the main sources of

the individual differences in rational thought that I have been exploring in this chapter. Such thinking dispositions vary systematically from individual to individual, and they are the source of what Meliorists consider the variance in the irrationalities in human cognition. Unlike the Panglossian (who assumes uniform rationality) or the Apologist (who minimizes such variability while not entirely denying it), the Meliorist is very accepting of the idea of variability in rational thought.

The Panglossian position in the Great Rationality Debate has obscured the existence of individual differences in rational thought and its underlying components. In particular, Panglossian philosophers have obscured the importance of the reflective mind. Philosophical treatments of rationality by Panglossians tend to have a common structure (see Cohen, 1981, for example). Such treatments tend to stress the importance of the competence/performance distinction and then proceed to allocate all of the truly important psychological mechanisms to the competence side of the dichotomy.

For example, Rescher (1988) argues that "to construe the data of these interesting experimental studies [of probabilistic reasoning] to mean that people are systematically programmed to fallacious processes of reasoning—rather than merely that they are inclined to a variety of (occasionally questionable) substantive suppositions—is a very questionable step.... While all (normal) people are to be credited with the capacity to reason, they frequently do not exercise it well" (p. 196). There are two parts to Rescher's (1988) point here: the "systematically programmed" part and the "inclination toward questionable suppositions" part (or, as Rips (1994, p. 394) puts it, whether incorrect reasoning is "systematically programmed or just a peccadillo"). Rescher's (1988) focus—like that of many who have dealt with the philosophical implications of the idea of human irrationality—is on the issue of how humans are "systematically programmed." "Inclinations toward questionable suppositions" are only of interest to those in the philosophical debates as mechanisms that allow one to drive a wedge between competence and performance—thus maintaining a theory of near-optimal human rational competence in the face of a host of responses that seemingly defy explanation in terms of standard normative models.

Analogously to Rescher, Cohen (1982) argues that there really are only two factors affecting performance on rational thinking tasks: "normatively

correct mechanisms on the one side, and adventitious causes of error on the other” (p. 252). Not surprisingly given such a conceptualization, the processes contributing to error (“adventitious causes”) are of little interest to Cohen (1981, 1982). In his view, human performance arises from an intrinsic human competence that is impeccably rational, but responses occasionally deviate from normative correctness due to inattention, memory lapses, lack of motivation, and other fluctuating but basically unimportant causes (in Cohen’s view). There is nothing in such a conception that would motivate any interest in patterns of errors or individual differences in such errors.

One of the goals of this chapter is to reverse the figure and ground in the rationality debate, which has tended to be dominated by the particular way that philosophers frame the competence/performance distinction. From a psychological standpoint, there may be important implications in precisely the aspects of performance that have been backgrounded in this controversy (“adventitious causes,” “peccadillos”). That is, whatever the outcome of the disputes about how humans are “systematically programmed,” variation in the “inclination toward questionable suppositions” is of psychological interest as a topic of study in its own right. The research discussed in this chapter provides at least tentative indications that the “inclination toward questionable suppositions” has some degree of domain generality and that it is predicted by thinking dispositions that concern the epistemic and pragmatic goals of the individual and that are part of the reflective mind.

Johnson-Laird and Byrne (1993; see Johnson-Laird, 2006) articulate a view of rational thought that parses the competence/performance distinction much differently from that of Cohen (1981, 1982, 1986). It is a view that highlights the importance of the reflective mind and leaves room for individual differences in important components of cognition. At the heart of the rational competence that Johnson-Laird and Byrne (1993) attribute to humans is not perfect rationality but instead just one meta-principle: People are programmed to accept inferences as valid provided that they have constructed no mental model of the premises that contradict the inference (see Johnson-Laird, Chapter 9). Inferences are categorized as false when a mental model is discovered that is contradictory. However, the search for contradictory models is “not governed by any systematic or comprehensive principles” (p. 178).

The key point in Johnson-Laird and Byrne’s (1993) account is that once an individual constructs a mental model from the premises, once the individual draws a new conclusion from the model, and once the individual begins the search for an alternative model of the premises that contradicts the conclusion, the individual “lacks any systematic method to make this search for counter-examples” (p. 205). Here is where Johnson-Laird and Byrne’s (1993) model could be modified to allow for the influence of thinking styles in ways that the impeccable competence view of Cohen (1981) does not. In this passage, Johnson-Laird and Byrne seem to be arguing that there are no systematic control features of the search process. But epistemically related thinking dispositions may in fact be reflecting just such control features.

Individual differences in the extensiveness of the search for contradictory models could arise from a variety of cognitive factors that, although they may not be completely systematic, may be far from “adventitious”—factors such as dispositions toward premature closure, cognitive confidence, reflectivity, dispositions toward confirmation bias, and ideational generativity. The decontextualizing requirement of many heuristics and biases tasks is a feature that is emphasized by many critics of that literature who, nevertheless, fail to see it as implying a research program for differential psychology. For example, I have argued that to contextualize a problem is such a ubiquitous reasoning style for human beings that it constitutes one of a very few so-called fundamental computational biases of information processing (Stanovich, 2003, 2004). Thus, it is not surprising that many people respond incorrectly when attempting a psychological task that is explicitly designed to require a decontextualized reasoning style (contrary-to-fact syllogisms, argument evaluation, etc.). But recall the empirically demonstrated variability on all of these tasks. The fact that some people *do* give the decontextualized response means that at least some people have available a larger repertoire of reasoning styles, allowing them to reason flexibly reason so as to override fundamental computational biases if the situation requires.

Another way of stressing the importance of individual differences in understanding the nature of rational thought is in terms of Dennett’s (1987) so-called intentional stance, which he marries to an assumption of idealized rationality. Dennett (1988) argues that we use the intentional stance for humans and dogs but not for lecterns because for the latter

“there is no predictive leverage gained by adopting the intentional stance” (p. 496). However, in several experiments discussed in this chapter, it has been shown that there is additional predictive leverage to be gained by relaxing the idealized rationality assumption of Dennett’s (1987, 1988) intentional stance and by positing measurable and systematic variation in intentional-level psychologies (that is, in the reflective mind). Knowledge about such individual differences in people’s intentional-level psychologies can be used to predict variance in the normative/descriptive gap displayed on many reasoning tasks. Consistent with the Meliorist conclusion that there can be individual differences in human rationality, the results show that there is variability in reasoning that cannot be accommodated within a model of perfect rational competence operating in the presence of performance errors and computational limitations.

Acknowledgments

This research was supported by grants from the Social Sciences and Humanities Research Council of Canada and the Canada Research Chairs program to Keith E. Stanovich. Laura Noveck and Keith Holyoak are thanked for their insightful comments on an earlier version of this chapter.

Notes

1. It should also be noted that in the view of rationality taken in this chapter, rationality is an intentional-level personal entity and not an algorithmic-level subpersonal one (Bermudez, 2001; Davies, 2000; Frankish, 2009; Stanovich, 1999, 2009a). A memory system in the human brain is not rational or irrational—it is merely efficient or inefficient (or of high or low capacity). Thus, subprocesses of the brain do not display rational or irrational properties per se, although they may contribute in one way or another to personal decisions or beliefs that could be characterized as such. Rationality concerns the actions of an entity in its environment that serve its goals. One of course could extrapolate the notion of environment to include the interior of the brain itself and then talk of a submodule that chose strategies rationally or not. This move creates two problems. First, what are the goals of this subpersonal entity—what are its interests that its rationality is trying to serve? This is unclear in the case of a subpersonal entity. Second, such a move regresses all the way down. We would need to talk of a neuron firing being either rational or irrational. As Oaksford and Chater (1998) put it, “the fact that a model is optimizing something does not mean that the model is a rational model. Optimality is not the same as rationality....Stomachs may be well or poorly adapted to their function (digestion), but they have no beliefs, desires or knowledge, and hence the question of their rationality does not arise” (pp. 4 and 5).

2. The principle of maximizing expected value says that the action that a rational person should choose is the one with the highest expected value. Expected value is calculated by taking the objective value of each outcome and multiplying it by the probability of that outcome and then summing those products over all of the possible outcomes. Symbolically, the formula is

as follows: Expected value = $\sum p_i v_i$; where p_i is the probability of each outcome and v_i is the value of each outcome. The symbol Σ is the summation sign, and simply means “add up all of the terms that follow.” The term *utility* refers to subjective value. Thus, the calculation of expected utility involves identical mathematics except that a subjective estimate of utility is substituted for the measure of objective value.

3. It is important to note that the Meliorist recognizes two different ways in which human decision-making performance might be improved. These might be termed *cognitive change* and *environmental change*. First, it might be possible to teach people better reasoning strategies and to have them learn rules of decision making that are helpful (see Stanovich, 2009b). These would represent instances of cognitive change. Additionally, however, research has shown that it is possible to change the environment so that natural human reasoning strategies will not lead to error (Gigerenzer, 2002; Milkman, Chugh, & Bazerman, 2009; Thaler & Sunstein, 2008). For example, choosing the right default values for a decision would be an example of an environmental change. In short, environmental alterations (as well as cognitive changes) can prevent rational thinking problems. Thus, in cases where teaching people the correct reasoning strategies might be difficult, it may well be easier to change the environment so that decision-making errors are less likely to occur.

4. Such empirical studies indicate that cognitive capacity and thinking dispositions measures are tapping separable variance. The converging evidence on the existence of this separable variance is growing (Bruine de Bruin, Parker, & Fischhoff, 2007; Finucane & Gullion, 2010; Klaczynski, Gordon, & Fauth, 1997; Klaczynski & Lavalée, 2005; Klaczynski & Robinson, 2000; Kokis et al., 2002; Macpherson & Stanovich, 2007; Newstead, Handley, Harley, Wright, & Farrelly, 2004; Parker & Fischhoff, 2005; Sá & Stanovich, 2001; Stanovich & West, 1997, 1998c, 2000; Toplak, Liu, Macpherson, Toneatto, & Stanovich, 2007; Toplak & Stanovich, 2002).

References

- Ackerman, P. L. (1994). Intelligence, attention, and learning: Maximal and typical performance. In D. K. Detterman (Ed.), *Current topics in human intelligence* (Vol. 4, pp. 1–27). Norwood, NJ: Ablex.
- Ackerman, P. L. (1996). A theory of adult development: Process, personality, interests, and knowledge. *Intelligence*, 22, 227–257.
- Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin*, 121, 219–245.
- Ackerman, P. L., & Kanfer, R. (2004). Cognitive, affective, and conative aspects of adult intellect within a typical and maximal performance framework. In D. Y. Dai & R. J. Sternberg (Eds.), *Motivation, emotion, and cognition: Integrative perspectives on intellectual functioning and development* (119–141). Mahwah, NJ: Lawrence Erlbaum Associates.
- Ainslie, G. (1982). A behavioral economic approach to the defence mechanisms: Freud’s energy theory revisited. *Social Science Information*, 21, 735–780.
- Ainslie, G. (2001). *Breakdown of will*. Cambridge, England: Cambridge University Press.
- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.
- Anderson, M. (2005). Marrying intelligence and cognition: A developmental view. In R. J. Sternberg & J. E. Pretz

- (Eds.), *Cognition and intelligence* (pp. 268–287). New York: Cambridge University Press.
- Arkes, H. R., & Ayton, P. (1999). The sunk cost and Concorde effects: Are humans less rational than lower animals? *Psychological Bulletin*, 125, 591–600.
- Aron, A. R. (2008). Progress in executive-function research: From tasks to functions to regions to networks. *Current Directions in Psychological Science*, 17, 124–129.
- Audi, R. (1993). *The structure of justification*. Cambridge, England: Cambridge University Press.
- Audi, R. (2001). *The architecture of reason: The structure and substance of rationality*. Oxford, England: Oxford University Press.
- Aunger, R., & Curtis, V. (2008). Kinds of behaviour. *Biology and Philosophy*, 23, 317–345.
- Austin, E. J., & Deary, I. J. (2002). Personality dispositions. In R. J. Sternberg (Ed.), *Why smart people can be so stupid* (pp. 187–211). New Haven, CT: Yale University Press.
- Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, 54, 462–479.
- Baron, J. (1982). Personality and intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 308–351). Cambridge, England: Cambridge University Press.
- Baron, J. (2008). *Thinking and deciding* (4th ed.). New York: Cambridge University Press.
- Barrett, H. C., & Kurzban, R. (2006). Modularity in cognition: Framing the debate. *Psychological Review*, 113, 628–647.
- Bates, T. C., & Shieles, A. (2003). Crystallized intelligence as a product of speed and drive for experience: the relationship of inspection time and openness to g and Gc. *Intelligence*, 31, 275–287.
- Bazerman, M., & Moore, D. A. (2008). *Judgment in managerial decision making*. New York: John Wiley.
- Bazerman, M., Tenbrunsel, A., & Wade-Benzoni, K. (1998). Negotiating with yourself and losing: Understanding and managing conflicting internal preferences. *Academy of Management Review*, 23, 225–241.
- Bermudez, J. L. (2001). Normativity and rationality in delusional psychiatric disorders. *Mind and Language*, 16, 457–493.
- Best, J. R., Miller, P. H., & Jones, L. L. (2009). Executive functions after age 5: Changes and correlates. *Developmental Review*, 29, 180–200.
- Bickerton, D. (1995). *Language and human behavior*. Seattle: University of Washington Press.
- Bishop, M. A., & Trout, J. D. (2005). *Epistemology and the psychology of human judgment*. Oxford, England: Oxford University Press.
- Brainerd, C. J., & Reyna, V. F. (2001). Fuzzy-trace theory: Dual processes in memory, reasoning, and cognitive neuroscience. In H. W. Reese & R. Kail (Eds.), *Advances in child development and behavior* (Vol. 28, pp. 41–100). San Diego, CA: Academic Press.
- Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology*, 92, 938–956.
- Cacioppo, J. T., Petty, R. E., Feinstein, J., & Jarvis, W. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, 119, 197–253.
- Camerer, C., Loewenstein, G., & Prelec, D. (2005). Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature*, 34, 9–64.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, England: Cambridge University Press.
- Carruthers, P. (2006). *The architecture of the mind*. New York: Oxford University Press.
- Cattell, R. B. (1963). Theory for fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54, 1–22.
- Cattell, R. B. (1998). Where is intelligence? Some answers from the triadic theory. In J. J. McArdle & R. W. Woodcock (Eds.), *Human cognitive abilities in theory and practice* (29–38). Mahwah, NJ: Erlbaum.
- Chaiken, S., Liberman, A., & Eagly, A. H. (1989). Heuristic and systematic information within and beyond the persuasion context. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 212–252). New York: Guilford Press.
- Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences*, 4, 317–370.
- Cohen, L. J. (1982). Are people programmed to commit fallacies? Further thoughts about the interpretation of experimental data on probability judgment. *Journal for the Theory of Social Behavior*, 12, 251–274.
- Cohen, L. J. (1986). *The dialogue of reason*. Oxford, England: Oxford University Press.
- Cokely, E. T., & Kelley, C. M. (2009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and Decision Making*, 4, 20–33.
- Coltheart, M. (1999). Modularity and cognition. *Trends in Cognitive Sciences*, 3, 115–120.
- Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, 58, 1–73.
- Cronbach, L. J. (1949). *Essentials of psychological testing*. New York: Harper.
- Davies, M. (2000). Interaction without reduction: The relationship between personal and sub-personal levels of description. *Mind and Society*, 1, 87–105.
- Dawes, R. M. (1998). Behavioral decision making and judgment. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (Vol. 1, pp. 497–548). Boston, MA: McGraw-Hill.
- Dawkins, R. (1982). *The extended phenotype*. New York: Oxford University Press.
- Del Missier, F., Mantyla, T., & Bruine de Bruin, W. (2010). Executive functions in decision making: An individual differences approach. *Thinking and Reasoning*, 16, 69–97.
- Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: The MIT Press.
- Dennett, D. C. (1988). Precis of “The Intentional Stance”. *Behavioral and Brain Sciences*, 11, 493–544.
- Dennett, D. C. (1995). *Darwin's dangerous idea: Evolution and the meanings of life*. New York: Simon & Schuster.
- Dennett, D. C. (1996). *Kinds of minds: Toward an understanding of consciousness*. New York: Basic Books.
- de Sousa, R. (2007). *Why think? Evolution and the rational mind*. Oxford, England: Oxford University Press.
- Doherty, M. (2003). Optimists, pessimists, and realists. In S. L. Schneider & J. Shanteau (Eds.), *Emerging perspectives on judgment and decision research* (pp. 643–679). New York: Cambridge University Press.

- Dohmen, T., Falk, A., Huffman, D., Marklein, F., & Sunde, U. (2009). Biased probability judgment: Evidence of incidence and relationship to economic outcomes from a representative sample. *Journal of Economic Behavior and Organization*, 72, 903–915.
- Duncan, J., Parr, A., Woolgar, A., Thompson, R., Bright, P., Cox, S.,...Nimmo-Smith, I. (2008). Goal neglect and Spearman's g: Competing parts of a complex task. *Journal of Experimental Psychology-General*, 137, 131–148.
- Edwards, W. (1954). The theory of decision making. *Psychological Bulletin*, 51, 380–417.
- Edwards, W., & von Winterfeldt, D. (1986). On cognitive illusions and their implications. In H. R. Arkes & K. R. Hammond (Eds.), *Judgment and decision making* (pp. 642–679). Cambridge, England: Cambridge University Press.
- Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, 11, 19–23.
- Evans, J. St. B. T. (1984). Heuristic and analytic processes in reasoning. *British Journal of Psychology*, 75, 451–468.
- Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences*. Hove, England: Erlbaum.
- Evans, J. St. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, 7, 454–459.
- Evans, J. St. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin and Review*, 13, 378–395.
- Evans, J. St. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgment*. New York: Psychology Press.
- Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment and social cognition. *Annual Review of Psychology*, 59, 255–278.
- Evans, J. St. B. T. (2009). How many dual-process theories do we need? One, two, or many? In J. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 33–54). Oxford, England: Oxford University Press.
- Evans, J. St. B. T. (2010). *Thinking twice: Two minds in one brain*. Oxford, England: Oxford University Press.
- Evans, J. St. B. T., & Frankish, K. (Eds.). (2009). *In two minds: Dual processes and beyond*. Oxford, England: Oxford University Press.
- Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction*. Hove, England: Erlbaum.
- Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, England: Psychology Press.
- Evans, J. St. B. T., & Over, D. E. (2010). Heuristic thinking and human intelligence: A commentary on Marewski, Gaissmaier and Gigerenzer. *Cognitive Processing*, 11, 171–175.
- Evans, J. St. B. T., & Wason, P. C. (1976). Rationalization in a reasoning task. *British Journal of Psychology*, 67, 479–486.
- Eysenck, H. J. (1994). Personality and intelligence: Psychometric and experimental approaches. In R. J. Sternberg & P. Ruzgis (Eds.), *Personality and intelligence* (pp. 3–31). Cambridge, England: Cambridge University Press.
- Finucane, M. L., & Gullion, C. M. (2010). Developing a tool for measuring the decision-making competence of older adults. *Psychology and Aging*, 25, 271–288.
- Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT University Press.
- Foley, R. (1987). *The theory of epistemic rationality*. Cambridge, MA: Harvard University Press.
- Frank, M. J., Cohen, M., & Sanfey, A. G. (2009). Multiple systems in decision making. *Current Directions in Psychological Science*, 18, 73–77.
- Frankish, K. (2004). *Mind and supermind*. Cambridge, England: Cambridge University Press.
- Frankish, K. (2009). Systems and levels: Dual-system theories and the personal-subpersonal distinction. In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 89–107). Oxford, England: Oxford University Press.
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19, 25–42.
- Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, 132, 692–731.
- Geary, D. C. (2005). *The origin of the mind: Evolution of brain, cognition, and general intelligence*. Washington, DC: American Psychological Association.
- Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review*, 103, 592–596.
- Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you*. New York: Simon & Schuster.
- Gigerenzer, G. (2007). *Gut feelings: The intelligence of the unconscious*. New York: Viking Penguin.
- Gilbert, D. T. (2006). *Stumbling on happiness*. New York: Alfred A. Knopf.
- Gilboa, I. (2010). *Rational choice*. Cambridge, MA: The MIT Press.
- Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. New York: Cambridge University Press.
- Goff, M., & Ackerman, P. L. (1992). Personality-intelligence relations: Assessment of typical intellectual engagement. *Journal of Educational Psychology*, 84, 537–552.
- Gray, J. R., Chabris, C. F., & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience*, 6, 316–322.
- Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108, 814–834.
- Harman, G. (1995). Rationality. In E. E. Smith & D. N. Osherson (Eds.), *Thinking* (Vol. 3, pp. 175–211). Cambridge, MA: The MIT Press.
- Hasher, L., Lustig, C., & Zacks, R. (2007). Inhibitory mechanisms and the control of attention. In A. Conway, C. Jarrold, M. Kane, A. Miyake, & J. Towse (Eds.), *Variation in working memory* (pp. 227–249). New York: Oxford University Press.
- Hastie, R., & Dawes, R. M. (2010). *Rational choice in an uncertain world* (2nd ed.). Thousand Oaks, CA: Sage.
- Hilton, D. J. (2003). Psychology and the financial markets: Applications to understanding and remedying irrational decision-making. In I. Brocas & J. D. Carrillo (Eds.), *The psychology of economic decisions. Vol. 1: Rationality and well-being* (pp. 273–297). Oxford, England: Oxford University Press.
- Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica*, 26, 1–23.
- Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In D. Flanagan, J. Genshaft, & P. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 53–91). New York: Guilford Press.

- Hsee, C. K., & Hastie, R. (2006). Decision and experience: Why don't we choose what makes us happy? *Trends in Cognitive Sciences*, 10, 31–37.
- Hurley, S., & Nudds, M. (2006). The questions of animal rationality: Theory and evidence. In S. Hurley & M. Nudds (Eds.), *Rational animals?* (pp. 1–83). Oxford, England: Oxford University Press.
- Jeffrey, R. C. (1983). *The logic of decision (Second Ed.)*. Chicago, IL: University of Chicago Press.
- Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.
- Johnson-Laird, P. N. (2006). *How we reason*. Oxford, England: Oxford University Press.
- Johnson-Laird, P. N., & Byrne, R. M. J. (1993). Models and deductive rationality. In K. Manktelow & D. Over (Eds.), *Rationality: Psychological and philosophical perspectives* (pp. 177–210). London: Routledge.
- Jungermann, H. (1986). The two camps on rationality. In H. R. Arkes & K. R. Hammond (Eds.), *Judgment and decision making* (pp. 627–641). Cambridge, England: Cambridge University Press.
- Kahneman, D. (2000). A psychological point of view: Violations of rational rules as a diagnostic of mental processes. *Behavioral and Brain Sciences*, 23, 681–683.
- Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, 58, 697–720.
- Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). New York: Cambridge University Press.
- Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 267–293). New York: Cambridge University Press.
- Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. *American Psychologist*, 64, 515–526.
- Kahneman, D., Krueger, A. B., Schkade, D., Schwarz, N., & Stone, A. (2006). Would you be happier if you were richer? A focusing illusion. *Science*, 312, 1908–1910.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430–454.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237–251.
- Kahneman, D., & Tversky, A. (1983). Can irrationality be intelligently discussed? *Behavioral and Brain Sciences*, 6, 509–510.
- Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, 103, 582–591.
- Kahneman, D., & Tversky, A. (Eds.). (2000). *Choices, values, and frames*. Cambridge, England: Cambridge University Press.
- Kanazawa, S. (2004). General intelligence as a domain-specific adaptation. *Psychological Review*, 111, 512–523.
- Kane, M. J. (2003). The intelligent brain in conflict. *Trends in Cognitive Sciences*, 7, 375–377.
- Kaufman, S. B., DeYoung, C. G., Gray, J. R., Jimenez, L., Brown, J., & Mackintosh, N. J. (2010). Implicit learning as an ability. *Cognition*, 116, 321–340.
- Klaczynski, P. A. (2001). Analytic and heuristic processing influences on adolescent reasoning and decision making. *Child Development*, 72, 844–861.
- Klaczynski, P. A., Gordon, D. H., & Fauth, J. (1997). Goal-oriented critical reasoning and individual differences in critical reasoning biases. *Journal of Educational Psychology*, 89, 470–485.
- Klaczynski, P. A., & Lavallee, K. L. (2005). Domain-specific identity, epistemic regulation, and intellectual ability as predictors of belief-based reasoning: A dual-process perspective. *Journal of Experimental Child Psychology*, 92, 1–24.
- Klaczynski, P. A., & Robinson, B. (2000). Personal theories, intellectual ability, and epistemological beliefs: Adult age differences in everyday reasoning tasks. *Psychology and Aging*, 15, 400–416.
- Koehler, D. J., & James, G. (2009). Probability matching in choice under uncertainty: Intuition versus deliberation. *Cognition*, 113, 123–127.
- Koehler, D. J., & James, G. (2010). Probability matching and strategy availability. *Memory and Cognition*, 38, 667–676.
- Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative and methodological challenges. *Behavioral and Brain Sciences*, 19, 1–53.
- Kokis, J., Macpherson, R., Toplak, M., West, R. F., & Stanovich, K. E. (2002). Heuristic and analytic processing: Age trends and associations with cognitive ability and cognitive styles. *Journal of Experimental Child Psychology*, 83, 26–52.
- Krueger, J., & Funder, D. C. (2004). Towards a balanced social psychology: Causes, consequences and cures for the problem-seeking approach to social cognition and behavior. *Behavioral and Brain Sciences*, 27, 313–376.
- Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing the mind: "Seizing" and "freezing". *Psychological Review*, 103, 263–283.
- Kuhberger, A. (2002). The rationality of risky decisions: A changing message. *Theory and Psychology*, 12, 427–452.
- Larrick, R. P. (2004). Debiasing. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 316–337). Malden, MA: Blackwell.
- Lee, C. J. (2006). Gricean charity: The Gricean turn in psychology. *Philosophy of the Social Sciences*, 36, 193–218.
- Lepine, R., Barrouillet, P., & Camos, V. (2005). What makes working memory spans so predictive of high-level cognition? *Psychonomic Bulletin and Review*, 12, 165–170.
- Leslie, A. M. (1987). Pretense and representation: The origins of "Theory of Mind". *Psychological Review*, 94, 412–426.
- Li, M., & Chapman, G. B. (2009). "100% of anything looks good:" The appeal of one hundred percent. *Psychonomic Bulletin and Review*, 16, 156–162.
- Lieberman, M. D. (2003). Reflexive and reflective judgment processes: A social cognitive neuroscience approach. In J. P. Forgas, K. R. Williams, & W. von Hippel (Eds.), *Social judgments: Implicit and explicit processes* (pp. 44–67). New York: Cambridge University Press.
- Lieberman, M. D. (2007). Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology*, 58, 259–289.
- Lieberman, M. D. (2009). What zombies can't do: A social cognitive neuroscience approach to the irreducibility of reflective consciousness. In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 293–316). Oxford, England: Oxford University Press.
- Loewenstein, G. F. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes*, 65, 272–292.

- Luce, R. D., & Raiffa, H. (1957). *Games and decisions*. New York: Wiley.
- Macpherson, R., & Stanovich, K. E. (2007). Cognitive ability, thinking dispositions, and instructional set as predictors of critical thinking. *Learning and Individual Differences*, 17, 115–127.
- Manktelow, K. I. (2004). Reasoning and rationality: The pure and the practical. In K. I. Manktelow & M. C. Chung (Eds.), *Psychology of reasoning: Theoretical and historical perspectives* (pp. 157–177). Hove, England: Psychology Press.
- Marewski, J. N., Gaissmaier, W., & Gigerenzer, G. (2010). Good judgments do not require complex cognition. *Current Directions in Psychological Science*, 11, 103–121.
- Matthews, G., Zeidner, M., & Roberts, R. D. (2002). *Emotional intelligence: Science and myth*. Cambridge, MA: MIT Press.
- McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science*, 306, 503–507.
- Metcalfe, J., & Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: Dynamics of will power. *Psychological Review*, 106, 3–19.
- Milkman, K. L., Chugh, D., & Bazerman, M. H. (2009). How can decision making be improved? *Perspectives on Psychological Science*, 4, 379–383.
- Minsky, M. L. (1985). *The society of mind*. New York: Simon and Schuster.
- Miyake, A., Friedman, N., Emerson, M. J., & Witzki, A. H. (2000). The utility and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*, 41, 49–100.
- Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, 132, 297–326.
- Murphy, D., & Stich, S. (2000). Darwin in the madhouse: Evolutionary psychology and the classification of mental disorders. In P. Carruthers & A. Chamberlain (Eds.), *Evolution and the human mind: Modularity, language and meta-cognition* (pp. 62–92). Cambridge, England: Cambridge University Press.
- Newstead, S. E., Handley, S. J., Harley, C., Wright, H., & Farrelly, D. (2004). Individual differences in deductive reasoning. *Quarterly Journal of Experimental Psychology*, 57A, 33–60.
- Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation* (pp. 1–18). New York: Plenum.
- Norris, S. P., & Ennis, R. H. (1989). *Evaluating critical thinking*. Pacific Grove, CA: Midwest Publications.
- Oaksford, M., & Chater, N. (1998). An introduction to rational models of cognition. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 1–18). New York: Oxford University Press.
- Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford, England: Oxford University Press.
- Oechssler, J., Roider, A., & Schmitz, P. W. (2009). Cognitive abilities and behavioral biases. *Journal of Economic Behavior and Organization*, 72, 147–152.
- Over, D. E. (2000). Ecological rationality and its heuristics. *Thinking and Reasoning*, 6, 182–192.
- Over, D. E. (2004). Rationality and the normative/descriptive distinction. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 3–18). Malden, MA: Blackwell.
- Parker, A. M., & Fischhoff, B. (2005). Decision-making competence: External validation through an individual differences approach. *Journal of Behavioral Decision Making*, 18, 1–27.
- Perkins, D. N. (1995). *Outsmarting IQ: The emerging science of learnable intelligence*. New York: Free Press.
- Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.
- Plato. (1945). *The republic*. (F. MacDonald Cornford, Trans.). New York: Oxford University Press.
- Pollock, J. L. (1995). *Cognitive carpentry*. Cambridge, MA: The MIT Press.
- Postman, N. (1988). *Conscientious objections*. New York: Vintage Books.
- Prado, J., & Noveck, I. A. (2007). Overcoming perceptual features in logical reasoning: A parametric functional magnetic resonance imaging study. *Journal of Cognitive Neuroscience*, 19, 642–657.
- Reber, A. S. (1993). *Implicit learning and tacit knowledge*. New York: Oxford University Press.
- Rescher, N. (1988). *Rationality: A philosophical inquiry into the nature and rationale of reason*. Oxford, England: Oxford University Press.
- Richerson, P. J., & Boyd, R. (2005). *Not by genes alone: How culture transformed human evolution*. Chicago, IL: University of Chicago Press.
- Rips, L. J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.
- Sá, W., Kelley, C., Ho, C., & Stanovich, K. E. (2005). Thinking about personal theories: Individual differences in the coordination of theory and evidence. *Personality and Individual Differences*, 38, 1149–1161.
- Sá, W., & Stanovich, K. E. (2001). The domain specificity and generality of mental contamination: Accuracy and projection in judgments of mental content. *British Journal of Psychology*, 92, 281–302.
- Sá, W., West, R. F., & Stanovich, K. E. (1999). The domain specificity and generality of belief bias: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology*, 91, 497–510.
- Salthouse, T. A., Atkinson, T. M., & Berish, D. E. (2003). Executive functioning as a potential mediator of age-related cognitive decline in normal adults. *Journal of Experimental Psychology: General*, 132, 566–594.
- Salthouse, T. A., & Pink, J. E. (2008). Why is working memory related to fluid intelligence? *Psychonomic Bulletin and Review*, 15, 364–371.
- Samuels, R. (2005). The complexity of cognition: Tractability arguments for massive modularity. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind* (pp. 107–121). Oxford, England: Oxford University Press.
- Samuels, R. (2009). The magical number two, plus or minus: Dual-process theory as a theory of cognitive kinds. In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 129–146). Oxford, England: Oxford University Press.
- Samuels, R., & Stich, S. P. (2004). Rationality and psychology. In A. R. Mele & P. Rawling (Eds.), *The Oxford handbook of rationality* (pp. 279–300). Oxford, England: Oxford University Press.
- Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.

- Schelling, T. C. (1984). *Choice and consequence: Perspectives of an errant economist*. Cambridge, MA: Harvard University Press.
- Schneider, W., & Chein, J. (2003). Controlled and automatic processing: Behavior, theory, and biological processing. *Cognitive Science*, 27, 525–559.
- Schommer-Aikins, M. (2004). Explaining the epistemological belief system: Introducing the embedded systemic model and coordinated research approach. *Educational Psychologist*, 39, 19–30.
- Shafir, E., & LeBoeuf, R. A. (2002). Rationality. *Annual Review of Psychology*, 53, 491–517.
- Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, 84, 127–190.
- Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119, 3–22.
- Sloman, S. A. (2002). Two systems of reasoning. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 379–396). New York: Cambridge University Press.
- Slovic, P., & Peters, E. (2006). Risk perception and affect. *Current Directions in Psychological Science*, 15, 322–325.
- Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review*, 4, 108–131.
- Sperber, D. (1994). The modularity of thought and the epidemiology of representations. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 39–67). Cambridge, England: Cambridge University Press.
- Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.
- Stanovich, K. E. (2001). Reductionism in the study of intelligence: Review of "Looking Down on Human Intelligence" by Ian Deary. *Trends in Cognitive Sciences*, 5(2), 91–92.
- Stanovich, K. E. (2002). Rationality, intelligence, and levels of analysis in cognitive science: Is dysrationalia possible? In R. J. Sternberg (Ed.), *Why smart people can be so stupid* (pp. 124–158). New Haven, CT: Yale University Press.
- Stanovich, K. E. (2003). The fundamental computational biases of human cognition: Heuristics that (sometimes) impair decision making and problem solving. In J. E. Davidson & R. J. Sternberg (Eds.), *The psychology of problem solving* (pp. 291–342). New York: Cambridge University Press.
- Stanovich, K. E. (2004). *The robot's rebellion: Finding meaning in the age of Darwin*. Chicago, IL: University of Chicago Press.
- Stanovich, K. E. (2009a). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 55–88). Oxford, England: Oxford University Press.
- Stanovich, K. E. (2009b). *What intelligence tests miss: The psychology of rational thought*. New Haven, CT: Yale University Press.
- Stanovich, K. E. (2010). *Decision making and rationality in the modern world*. New York: Oxford University Press.
- Stanovich, K. E. (2011). *Rationality and the reflective mind*. New York: Oxford University Press.
- Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology*, 89, 342–357.
- Stanovich, K. E., & West, R. F. (1998a). Cognitive ability and variation in selection task performance. *Thinking and Reasoning*, 4, 193–230.
- Stanovich, K. E., & West, R. F. (1998b). Individual differences in framing and conjunction effects. *Thinking and Reasoning*, 4, 289–317.
- Stanovich, K. E., & West, R. F. (1998c). Individual differences in rational thought. *Journal of Experimental Psychology: General*, 127, 161–188.
- Stanovich, K. E., & West, R. F. (1998d). Who uses base rates and $P(D/-H)$? An analysis of individual differences. *Memory and Cognition*, 26, 161–179.
- Stanovich, K. E., & West, R. F. (1999). Discrepancies between normative and descriptive models of decision making and the understanding/acceptance principle. *Cognitive Psychology*, 38, 349–385.
- Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23, 645–726.
- Stanovich, K. E., & West, R. F. (2003). Evolutionary versus instrumental goals: How evolutionary psychology misconceives human rationality. In D. E. Over (Ed.), *Evolution and the psychology of thinking: The debate* (pp. 171–230). Hove, England and New York: Psychology Press.
- Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking and Reasoning*, 13, 225–247.
- Stanovich, K. E., & West, R. F. (2008a). On the failure of intelligence to predict myside bias and one-sided bias. *Thinking and Reasoning*, 14, 129–167.
- Stanovich, K. E., & West, R. F. (2008b). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, 94, 672–695.
- Stein, E. (1996). *Without good reason: The rationality debate in philosophy and cognitive science*. Oxford, England: Oxford University Press.
- Sternberg, R. J. (2003). *Wisdom, intelligence, and creativity synthesized*. Cambridge, England: Cambridge University Press.
- Sternberg, R. J., & Grigorenko, E. L. (1997). Are cognitive styles still in style? *American Psychologist*, 52, 700–712.
- Sternberg, R. J., Grigorenko, E. L., & Zhang, L. (2008). Styles of learning and thinking matter in instruction and assessment. *Perspectives on Psychological Science*, 3, 486–506.
- Stich, S. P. (1990). *The fragmentation of reason*. Cambridge, MA: MIT Press.
- Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, 8, 220–247.
- Strathman, A., Gleicher, F., Boninger, D. S., & Scott Edwards, C. (1994). The consideration of future consequences: Weighing immediate and distant outcomes of behavior. *Journal of Personality and Social Psychology*, 66, 742–752.
- Taub, G. E., & McGrew, K. S. (2004). A confirmatory factor analysis of Cattell-Horn-Carroll theory and cross-age invariance of the Woodcock-Johnson Tests of Cognitive Abilities III. *School Psychology Quarterly*, 19, 72–87.

- Terlock, P. E., & Mellers, B. A. (2002). The great rationality debate. *Psychological Science*, 13, 94–99.
- Thagard, P., & Nisbett, R. E. (1983). Rationality and charity. *Philosophy of Science*, 50, 250–267.
- Thaler, R. H., & Shefrin, H. M. (1981). An economic theory of self-control. *Journal of Political Economy*, 89, 392–406.
- Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven, CT: Yale University Press.
- Toates, F. (2005). Evolutionary psychology: Towards a more integrative model. *Biology and Philosophy*, 20, 305–328.
- Toates, F. (2006). A model of the hierarchy of behavior, cognition, and consciousness. *Consciousness and Cognition*, 15, 75–118.
- Toplak, M., Liu, E., Macpherson, R., Toneatto, T., & Stanovich, K. E. (2007). The reasoning skills and thinking dispositions of problem gamblers: A dual-process taxonomy. *Journal of Behavioral Decision Making*, 20, 103–124.
- Toplak, M. E., & Stanovich, K. E. (2002). The domain specificity and generality of disjunctive reasoning: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology*, 94, 197–209.
- Toplak, M. E., & Stanovich, K. E. (2003). Associations between myside bias on an informal reasoning task and amount of post-secondary education. *Applied Cognitive Psychology*, 17, 851–860.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131.
- von Neumann, J., & Morgenstern, O. (1944). *The theory of games and economic behavior*. Princeton, NJ: Princeton University Press.
- Vranas, P. B. M. (2000). Gigerenzer's normative critique of Kahneman and Tversky. *Cognition*, 76, 179–193.
- Wang, L. (2009). Money and fame: Vividness effects in the National Basketball Association. *Journal of Behavioral Decision Making*, 22, 20–44.
- Wason, P. C., & Evans, J. St. B. T. (1975). Dual processes in reasoning? *Cognition*, 3, 141–154.
- West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology*, 100, 930–941.
- Wilson, T. D. (2002). *Strangers to ourselves*. Cambridge, MA: Harvard University Press.
- Wilson, T. D., & Gilbert, D. T. (2005). Affective forecasting: Knowing what to want. *Current Directions in Psychological Science*, 14, 131–134.
- Wu, G., Zhang, J., & Gonzalez, R. (2004). Decision under risk. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 399–423). Malden, MA: Blackwell Publishing.
- Zeidner, M., & Matthews, G. (2000). Intelligence and personality. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 581–610). New York: Cambridge University Press.
- Zelazo, P. D. (2004). The development of conscious control in childhood. *Trends in Cognitive Sciences*, 8, 12–17.

Cognition and the Creation of Ideas

Steven M. Smith and Thomas B. Ward

Abstract

The cognition that gives rise to creative thinking is not a singular process or operation; rather, it consists of many different cognitive structures and processes that can collaborate in a variety of ways to construct different types of creative products. Cognition that is often relevant to creativity includes remote association, conceptual combination, visualization, retrieving and mapping analogies, reasoning, insight problem solving, implicit and explicit cognition, abstraction, and mental models.

Key Words: creativity, cognition, insight, imagination, emergence

How do our minds produce creative ideas? Clearly, there can be no formulaic answer to this enormous question. Part of the problem is that a consensual definition of creativity is difficult to construct. There are many domains of creative endeavor, such as art, engineering, literature, business, or athletics, and creativity might systematically differ for different domains. Furthermore, even within a single domain there are usually many different ways that creative ideas come into being. Even a single individual can be creative in different ways at different times. To know how our minds produce creative ideas, we undoubtedly need to understand the many factors that affect creativity, and there are multiple approaches to understanding those factors. For example, motivational factors, emotional factors, individual differences, environmental factors, and historical factors can strongly influence creativity. One of those factors, clearly, must be cognition: Creativity is at least in part influenced or even determined by cognition. In this chapter we focus on the question of how cognition supports or gives rise to creative work—the *creative cognition* approach.

The creative cognition approach begins with the rather obvious assumption that cognitive processes

play a critical role when people get creative ideas. Creative ideas are ones that are novel and potentially of value. There are similarities among cases in which people have had historically influential ideas, or ideas that are consensually judged as creative. Research in creative cognition is based upon the premise that similar experiences can be evoked in laboratory experiments. Metrics of creativity that emerge from these studies measure aspects of creativity but generally not how creative the ultimate product is. Most who use this creative cognition approach endorse the notion that there is no singular cognitive process or mental operation that we could call *the creative process*. Rather, we manage our cognitive resources in various ways to try to produce ideas with novelty and potential value, and the way those cognitive resources are managed may differ for different domains, different individuals, and different situations. In the course of pursuing creative ideas, people often experience certain phenomena or use certain heuristics to make creative ideas more likely.

We also endorse the notion that the cognition involved in creative activities resembles the cognition

that operates in other areas of cognition that are not commonly considered creative. For example, false memories are created by the same processes that affect creative ideation, analogies are mapped in creative and noncreative cognition, we create language when we speak, and we create mental models of our environments. Novel and useful outcomes emerge in many domains of cognition. What is special about creative cognition is that the insights and ideas produced are usually unexpected (e.g., Metcalfe, 1986).

The less surprising the product, the less creative it seems to be; nonobviousness is a requirement of all U.S. patents (Smith, 2008). Furthermore, unexpected insights and ideas that have value prompt the experience of delight. Although it engages the same basic cognitive mechanisms involved in domains such as memory, concept formation, problem solving, decision making, and metacognition, creative cognition is not limited to generating planned or deliberately intended products. Truly creative cognition can bring into being unexpected products of delight.

Domains of Cognition Involved in Creative Thinking

There are multiple domains of cognition that are involved in creative thinking, essentially, the same that are involved in noncreative cognition. The most noteworthy of these domains include conceptualization, visualization, memory, problem solving, language, decision making, and several areas of implicit cognition.

Concepts and Categories: The Structures of Imagination

Ward's (1994) structured imagination theory centers on the notion that the ideas imagined by creating minds are based, at least in part, on the conceptual cognitive structures that must be used, extended, and combined during the course of creative cognition. Understanding the ways that concepts are flexibly constructed and tailored for different contexts, how they are combined, and what guides the way that people extend categories in order to recognize or generate new instances of those concepts is essential to this structured imagination view.

Humans acquire vast storehouses of knowledge as a result of their experiences in the world. Because of the centrality of this type of knowledge to what it means to be human, considerable research has been devoted to the nature and structure of our concepts and categories, how they are acquired, and how they guide our actions and thought (Barsalou,

1985; Murphy & Medin, 1985; Rosch, 1973, 1975; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; see Rips et al., Chapter 11). Rather than being randomly organized collections of separate pieces of information, concept representations are highly organized and, at least from a functional perspective, hierarchical (e.g., Bower, Clark, Winzenz, & Lesgold, 1969). Both of these properties have implications for understanding the role of cognition in creativity.

Before considering the structured and hierarchical aspects of concepts, it is worth noting that the formation of concepts is itself a kind of creative, or at the very least generative, activity. This is because experiences with entities in the world are discrete, and the structures that tie those experiences together must be generated by the learner. Consider the child who ultimately learns the category "dog" from experiences with the family pet, others seen from the window, in the park, in books, or other places. Each of the experiences occurs at a separate time point, and the "glue" holding them together must be generated by the child. At a more extreme level, even the conclusion that the family pet is really the same entity across multiple discrete exposures is a kind of construction, not given directly by the input. Although this type of construction of conceptual structures is much simpler than those typically thought of when considering creativity, views of creativity are broadening to include a wider range of phenomenon. For example, Kaufman and Beghetto (2009) introduced the *four C* model of creativity to deal with the perceived inadequacy of early suggestions regarding the split between small-c (everyday) and Big-C (eminent) creativity. The latter split does not really allow a distinction between, for example, the songwriting of an amateur amusing himself with simple creations and a professional making her living writing songs, both of which would be small-c creativity, in contrast to the lasting, Big-C works of acknowledged master composers, such as Mozart. Nor does the small-c versus Big-C split allow the type of personally constructed understandings that are developed by individuals as they attempt to make sense of their world to be thought of as creativity, because they may not result in observable products that a qualified audience would judge to be novel and useful. Yet those understandings are clearly an instance of generative thinking. To deal with the latter problem, Kaufman and Beghetto (2009, see also Beghetto & Kaufman, 2007) introduced mini-c creativity to capture the

notion that such personal understandings are a kind of creative construction. By extension, the conceptual structures people generate in service of organizing and understanding experiences can be thought of as instances of creativity, albeit at a very simple level. Thus, Kaufman's four Cs include mini-c (personal realizations), small-c (everyday insights), pro-c (advanced enough to be making a career of some creative domain), and Big-C (eminent contributions to chosen domain).

The fact that our conceptual knowledge is organized and structured into identifiable groupings also relates to another aspect of creativity, namely, the degree to which it is domain general versus domain specific (e.g., Finke et al., 1992). This issue is concerned with the question of whether creative capacity is better thought of as a general ability that could be applied to a wide range of domains or a more specialized ability that would facilitate performance in a single domain. Whether people possess a more general or more specific creativity capacity, the fact is that they create within domains that can be described at various levels of breadth, such as written communication, fiction writing, novel writing, romance novel writing, and so on. To the extent that people possess differing amounts of knowledge and differently structured knowledge about a given domain, a consideration of conceptual structure would seem to favor a domain-specific view of creativity. Certainly, as noted later, domain knowledge will influence the form of newly created ideas with a domain. Alternatively, since some of the basic processes by which concepts are retrieved, combined, and used might represent general cognitive tendencies, a focus on conceptual processing could be readily compatible with a more domain-general view of creativity.

Our organized knowledge structures allow us to function effectively in the world and communicate with others. At a most fundamental level they allow us to identify and correctly classify new entities rather than have to treat each newly encountered object as something novel to be learned about (see Rips et al., Chapter 11). We can readily classify a furry, four-legged, barking, tail-wagging creature with floppy ears dangling from the sides of its head as a dog even if we have never seen it or even another of its breed before because of the stored information we have acquired about previously encountered dogs. Although this capacity for rapid, efficient classification serves us well in most circumstances, it can also underlie our tendency to miss potentially

creative solutions to even the simplest of problems. One classic example of this is the two-string problem, in which would-be solvers must find a way to tie together two strings suspended from the ceiling that are too far apart to be grasped simultaneously (Maier, 1931). Even though a pair of pliers is present that could serve as a pendulum weight for swinging the distant string closer to the one the solver is grasping, most do not readily see that possibility, due at least in part to the efficient tendency to categorize them as pliers rather than as a heavy object. Glucksberg and Weisberg (1966) showed that this so-called functional fixedness effect was prevented somewhat if participants verbally labeled the critical incidental stimulus, adding evidence that categorization can powerfully impact fixation.

The hierarchical aspect of conceptual structure is also important in considering creativity from a cognitive perspective. To illustrate, the same entity could be thought of as a tabby, a cat, a feline, a mammal, a living thing, and so on. Classic research established that, rather than all levels being equally prominent, the basic level, intermediate between the most general and the most specific, tends to predominate in initial classification, labeling and reasoning about a given entity (Rosch, 1973; Rosch et al., 1976; Rosch, 1975). In the example given here, "cat" would be the basic level. The importance of this phenomenon for creativity is that, as described more fully in a later section, the basic level also plays a powerful role in the generation of new ideas within a domain (Rosch, 1973; Rosch et al., 1976; Rosch, 1975).

Our concepts also allow us to go beyond simple classification and to reason and draw inferences about known and novel category instances (Gelman, 1988; Heit, 2000; Osherson, Smith, Wilkie, López & Shafir, 1990; Rips, 1975). Conceptual knowledge includes considerably more information than just that which might be used to classify objects. It also includes known or assumed properties that we can generalize to new instances. Even without seeing the entity described in the previous paragraph, on being told that it is a dog, we would make certain inferences with varying degrees of certainty, including the likelihood that it possessed those identifying properties, as well as a number of others including that it might have a heart, lungs, or a stomach, be someone's pet, go for walks in the park with its owner, eat and drink from bowls, and be susceptible to fleas and heartworms. The extent to which each of those inferences is warranted could be debated, and

considerable research has been devoted to understanding the factors that influence the tendency of people to draw inferences (e.g., Heit, 2000). But such tendencies are also direct determinants of the novelty or unusualness of newly generated products. As noted earlier, people tend to create within domains, and they show a striking tendency to incorporate properties of known domain instances into the novel items they generate (e.g., Bredart, Ward, & Marczewski, 1998; Ward, 1994; Ward, Dodds, Saunders, & Sifonis, 2000; Ward, Patterson, Sifonis, Dodds, & Saunders, 2002) (see Fig. 23.1).

According to the path-of-least-resistance theory (Ward, 1994, 1995), when people develop new ideas, such as designing imaginary life forms, they tend to begin by retrieving highly representative exemplars of creatures, such as a dog, a fish, or a bird. Next, they project the properties of those specific instances onto the novel ideas they are developing. Taking this path of least resistance leads to less original ideas and thwarts more flexible uses of conceptual knowledge. Abstraction can help one avoid this reliance on representative instances, making one more likely to produce creative ideas (Ward & Sifonis, 1997).

Memory

The link between memory and creativity may seem counterintuitive, because memory seems designed to converge upon real events of the past, whereas creativity seems more divergent, dealing with imaginative possibilities. In fact, memory and creativity have much in common, even if the products of memory processes are often quite different from the products of creative cognition. Both creativity and memory involve both conscious and implicit processes, both are sensitive to recent and frequent experiences, both show blocking and recovery effects, and both involve constructive cognitive processes.

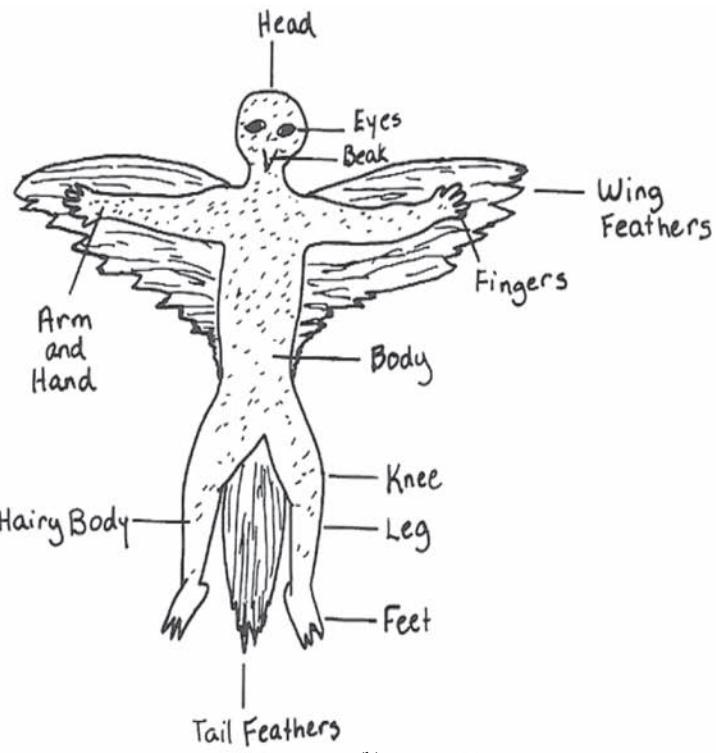
Implicit memory is a cognitive resource that supports conscious operations in many ways (see Evans, Chapter 8). For example, in automatization, a repetitious set of operations is offloaded from the explicit system to an automatic, implicit set of processes; this offloading frees up cognitive resources for other attention-demanding activities. Unfortunately, when implicit memory makes inappropriate material too accessible, it can block access to more appropriate information. Worse, this type of block happens without conscious awareness, so it is difficult for one to detect one's own implicit memory blocks. Smith

and Tindell (1997) showed that reading a word (e.g., ANALOGY) that is orthographically similar to the correct completion of a word fragment (e.g., A_L__GY) can cause an implicit memory block for the solution (ALLERGY) (see Fig. 23.2).

Participants who were warned that previously seen blocker words were incorrect solutions still remain susceptible to these implicit memory blocks. Similarly, in creative invention tasks, designers automatically incorporate features of previously seen examples, even when those features are problematic and explicitly forbidden in the instructions (Jansson & Smith, 1991; Landau & Lehr, 2004; Smith, Ward, & Schumacher, 1993). Only when the specific reasons that negative features are problematic are clearly explained to participants is this design fixation effect mitigated (Chrysikou & Weisberg, 2005). Thus, implicit memory of inappropriate material appears to block creative design (see Figs. 23.3 and 23.4).

Another area in which memory and creative cognition share a great deal of overlap involves recovery from blocks. Two mysterious phenomena, *reminiscence* (and the related phenomenon *hypermnesia*) in memory, and *incubation* effects in creative problem solving, have been linked to the same underlying cognitive mechanisms. Reminiscence refers to a phenomenon in which memories that are initially inaccessible are subsequently retrieved without re-exposure to the to-be-remembered material. Hypermnesia, a closely related phenomenon, is a net increase in recall from one recall attempt to the next (e.g., Erdelyi & Becker, 1974). These memory phenomena fly in the face of the long-established rule that memory gets *worse* as the retention interval increases, because time elapses from one test to the next. The mystery of incubation effects is that key ideas for difficult problems sometimes occur when one takes a break from the problem, rather than working on it uninterrupted (e.g., Smith & Blankenship, 1989, 1991). These key ideas are not predictable by the people who experience the moments of insight (e.g., Metcalfe & Weibe, 1987), and they may result from cognitive operations that operate outside of awareness (e.g., Schooler & Melcher, 1995).

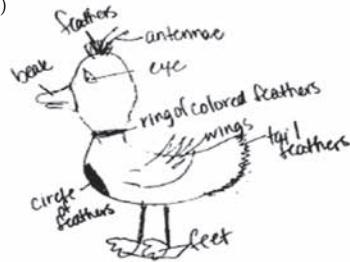
Although there are multiple explanations of reminiscence and of incubation effects, and these phenomena may be multiply caused, a single theory has been proposed to explain both. The *forgetting fixation* theory (e.g., Kohn & Smith, 2009; Smith & Blankenship, 1989, 1991) states that both reminiscence and incubation effects can be caused by recovery from initial blocks (i.e., fixation) when



(a)



(b)



(c)



(d)

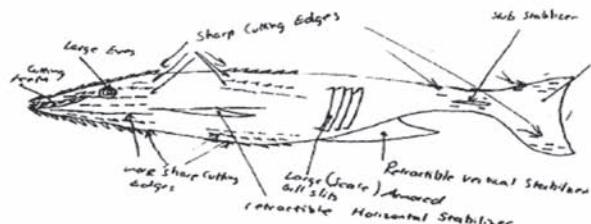


Fig. 23.1 At the top, a novel imaginary creature designed by a participant includes features of known creatures (e.g., head with sense organs, appendages, bilateral symmetry). The four sketches at the bottom are from conditions in which participants were told to create (a) any creature from an imaginary planet like Earth, (b) one with feathers, (c) one with fur, and (d) an imaginary creature with scales (Ward et al., 2002). Created ideas also tend to include properties generalized from known creatures with feathers (birds), fur (mammals), or scales (fish).

<u>Blocker</u>	<u>Fragment</u>	<u>Solution</u>
ANALOGY	A_L_GY	ALLERGY
BRIGADE	B_G_A_E	BAGGAGE
COTTAGE	C_TA_G	CATALOG
CHARTER	CHAR_T_	CHARITY
CLUSTER	C_U_TR_	COUNTRY
CRUMPET	CU_P_T	CULPRIT
DENSITY	D_NITY	DIGNITY
FIXTURE	F_I_URE	FAILURE
HOLSTER	H_ST_R_	HISTORY
TONIGHT	T_NG_T	TANGENT
TRILOGY	TR_G_Y	TRAGEDY
VOYAGER	VO_ AGE	VOLTAGE

Fig 23.2 Word fragments, their correct solutions, and blocker words used to show an implicit memory blocking effect (Smith & Tindell, 1997).

blocking responses are put out of mind. In the case of incubation effects, many puzzles or insight problems initially may promote ideas or ways of thinking that are ultimately found to be inappropriate solutions—blockers—for those problems. The forgetting fixation hypothesis states that getting blockers out of mind can benefit creative problem solving. The same hypothesis explains reminiscence by focusing on *output interference* that occurs on the initial recall test; as items are recalled from a memory set, their subsequent probability of being retrieved again is increased. This results in a biased retrieval set in which initially recalled (and strengthened) items block recall of other items in the same memory set. The forgetting fixation hypothesis explains reminiscence as a decrease (after a time lag) in the blocking effect exerted by initially retrieved items in the same memory set, thereby allowing initially blocked items to be retrieved (e.g., Smith, 1995a).

Empirical tests of the forgetting fixation hypothesis in reminiscence and incubation show clear support for the hypothesis. Reminiscence, defined as recalling on a second test items not recalled on an earlier test, can be explained by the incremental hypothesis as the result of continued retrieval efforts, and since more attempts will eventually find initially unrecalled items, those items might be recalled on a second test (Roediger & Thorpe, 1978). What if a second (unexpected) recall test is given after a delay during which participants are kept busy? The incremental hypothesis predicts that reminiscence should drop if the retest is delayed because initially unretrieved items are even harder to retrieve after a delay. The forgetting fixation hypothesis, however, predicts that a delay permits

weakening of initial output interference, resulting in greater reminiscence on a delayed test. Smith and Vela (1991) showed *incubated reminiscence* effects, consistent with the forgetting fixation prediction: More initially unretrieved items were recalled if a retest was delayed.

A similar pattern of results has been found in creative problem solving, using puzzle problems (Smith & Blankenship, 1989) and Remote Associates Test problems (Kohn & Smith, 2009; Smith & Blankenship, 1991; Vul & Pashler, 2007) (see Fig 23.5). To ensure that initial fixation occurs, some participants in these studies were exposed to misleading hints before their first attempt at solving these problems, so initial fixation was experimentally introduced, and its effect verified by poorer initial performance relative to control group conditions. Retesting immediately versus after a delay yields consistent incubation effects; that is, problems not initially solved are more likely to be solved after a delay, as compared to immediate retesting conditions. Furthermore, these incubation effects are more robust when fixation is experimentally introduced. Thus, the forgetting fixation hypothesis, which explains reminiscence effects, also explains incubation effects in creative problem solving (Fig 23.6).

Incubation effects in tip-of-the-tongue (TOT) resolution have also been reported. When a name or word that could not be retrieved initially is later remembered, it often seems to happen in the absence of the initial TOT experience. Because TOT states have sometimes been attributed to momentary memory blocks (e.g., Jones, 1989; Jones & Langford, 1987; Reason & Lucas, 1983), these TOT recovery experiences may reflect the same underlying cognitive processes that give rise to incubation effects. Choi and Smith (2005) asked participants to name eight capital cities, each cued by the names of the country, eight names of diseases cued by disease descriptions, eight names of celebrities cued by their photographs, and so on, asking for graded TOT judgments whenever a name or word could not be remembered. Resolution of TOT states (i.e., successful recall of initially unrealled items) was greater if retesting followed an incubation period than if retesting was not delayed, and this TOT incubation effect was greater for TOT states that had been judged stronger. It was assumed that some of these TOTs were caused by blocking because not one, but eight questions were given for each category, a procedure known to cause output interference (e.g., Brown & Hall, 1979). On the other hand, Kornell and Metcalfe (2006), who

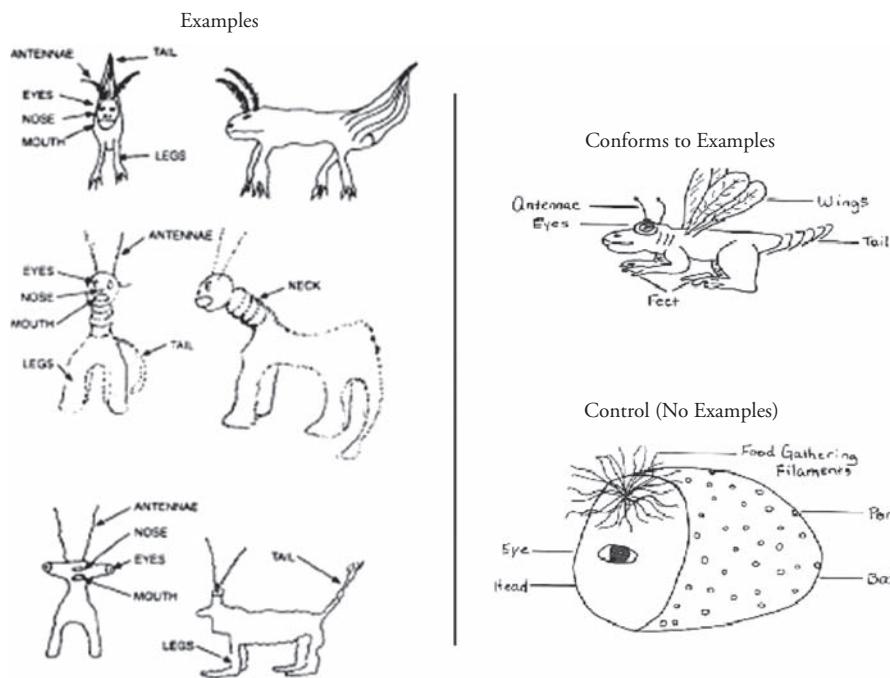


Fig. 23.3 The left panel shows examples seen by half the participants, who all tried to create life forms that might evolve on an imaginary planet similar to Earth. The top right shows an idea from a participant who saw the three examples; this sketch contains all three of the exemplified features (four legs, antennae, tail), a conformity effect that is found even when participants are asked to design ideas that are as different as possible from the examples. The bottom right shows an idea with none of those features, sketched by a participant who saw no examples (Smith et al., 1993).

also found incubation effects in TOT resolution, found that incubation did not interact with initial blocking self-reports, and reminding participants of their reported blockers at retest did not affect TOT resolution, casting doubt on the forgetting fixation hypothesis of TOT resolution. In that study, blocking was not directly manipulated but rather defined in terms of metacognitive reports of blocker experiences. Thus, whether incubated TOT resolution can be explained by forgetting blockers remains in question, probably due to the likelihood that TOTs have multiple causes, such as partial activation, memory blocking, and experiment demand characteristics.

Problem Solving

Creative and noncreative problem solving can be distinguished. Using known algorithms or heuristics to solve a problem, or simply retrieving a known solution from memory, would be considered noncreative problem solving, whereas creative problem solving involves solving new problems or solving old problems in new ways. Not all creative thinking can be described as creative problem solving, but much creative cognition research frames the activity that way.

One distinction that separates creative and noncreative problem solving is the concept of *well-defined*

versus *ill-defined* problems. A problem is considered well defined (and not creative) if its beginning state and goal state are thoroughly specified, in addition to the operations to be used in getting from the beginning to the goal (Reitman, 1965). Creative problems are considered ill defined, primarily because multiple hypothetical solutions might satisfy the goals of the problem. Another approach distinguishes *divergent* from *convergent problem solving*. Whereas a convergent problem has a single correct answer, divergent problems have many possible solutions (e.g., Finke, Ward, & Smith, 1992). Divergent problem solving will be discussed in the next section.

The topic of insight (e.g., Dominowski & Jenrick, 1972; see also van Steenburgh et al., Chapter 24) often has been studied as a type of creative problem solving. It is important to distinguish the concepts of insight, insight problems, and insight experiences. Whereas *insight* refers to a clear and/or deep understanding, *insight experiences* (also known as *aha!* experiences) are insights that are experienced suddenly and unpredictably. *Insight problems* are puzzle problems that are typically (but not necessarily) solved by insight experiences (Finke et al., 1992; Smith, 1995a). Insight problem solving, as discussed earlier, often has been described as accompanied by a perceptual-like restructuring of

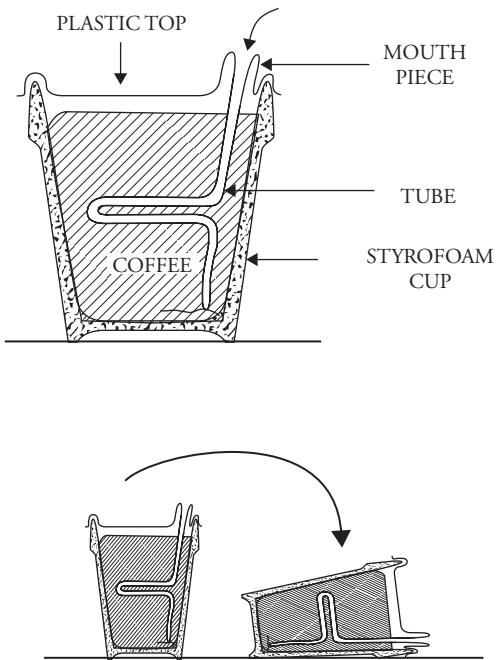


Fig. 23.4 Example spill-proof coffee cup given to some participants in Jansson and Smith (1991). Although the example uses a straw and a mouthpiece (which could lead to scalded mouths), these features were explicitly disallowed for designs sketched by participants. Nonetheless, design students who saw this example were far more likely to include straws and mouthpieces in the designs they sketched, as compared with students who did not see the example.

the problem, that is, discovering a new or different way to frame a problem in a manner similar to the way that an alternative interpretation of an ambiguous figure might be discovered (e.g., Dominowski, 1981; Duncker, 1945; Ellen, 1982; Maier, 1930; Metcalfe, 1986) (see Fig. 23.7). Initial attempts to solve insight problems often result in fixation, impediments caused by retrieval of inappropriate prepotent responses or methods. Recent research suggests that inhibition of prepotent responses may facilitate insight problem solving by providing a means by which to overcome fixation (Storm & Angello, 2010).

Implicit Cognition

Cognitive processes that occur without explicit awareness or deliberate intention can be described as *implicit cognition*. Implicit or unconscious cognition has long been considered to play a critically important role in the creative process. The idea that insights or solutions to problems can be arrived at via unconscious processes has been long endorsed as the reason that creative solutions seem to appear unbidden, and often occur outside of the typical problem-solving context, even when conscious work has been put aside. On the one hand, no evidence indicates that unconscious cognitive processes, like conscious processes, are capable autonomously of carrying out sequences of knowledge states, with each step using the products of previous states. On the other hand, there has been evidence that unconscious processes similar to spreading activation might be responsible for intuitive hunches (e.g., Bowers, Farvolden, & Mermigis, 1995), that impending solutions to insight problems cannot be predicted more in advance than about 10 seconds (Metcalfe, 1986), and that inventors and students in creative fields are often subject to involuntary conformity effects (e.g., Smith, Ward, & Schumacher, 1993).

The type of implicit cognition most often referred to in reference to creative thinking is spreading activation that occurs below the level of conscious awareness. Below-threshold activation has been the mechanism for a theory that implicit cognition can lead incrementally, but below the level of conscious awareness, toward an insight, and the final step in the process that completes an idea is simply the product of numerous incremental steps, not a cognitive restructuring (e.g., Weisberg, 1992). Yaniv and Meyer (1987) found that answers to general knowledge questions were responded to faster in a lexical decision task even if a participant could not consciously retrieve the answer, and furthermore, the stronger the feeling-of-knowing report accompanying the initial retrieval failure, the clearer the lexical decision advantage. They attributed this

Remote Associates Test Problems			Blockers	Solutions
SALAD	HEAD	GOOSE	lettuce	egg
BED	DUSTER	WEIGHT	room	feather
APPLE	HOUSE	FAMILY	green	tree
CAT	SLEEP	BOARD	black	walk
WATER	SKATE	CUBE	sugar	ice
ARM	COAL	STOP	rest	pit

Fig. 23.5 Remote Associates Test (RAT) problems, along with examples of blockers (related to two but not all three test words) and solutions (a single word related to each of the three test words).

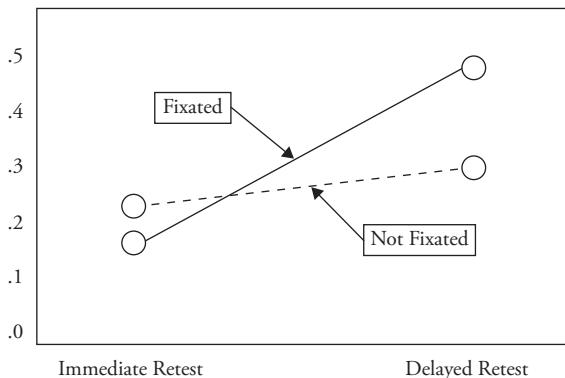


Fig. 23.6 Proportion of initially failed problems that were resolved on a retest that was given either immediately, or after a 5-minute delay. Problems with no misleading “hints” on the original test (Not Fixated) did not show an incubation effect (defined as better resolution on the delayed retest), but those with misleading hints (Fixated) showed a clear incubation effect (from Experiment 1 of Smith & Blankenship, 1991).

priming effect to below-threshold activation of target items, caused by the unsuccessful retrieval attempts. Although Connor, Balota, and Neely (1992) later challenged the interpretation of their results, Yaniv and Meyer claimed that such activation could be the mechanism underlying incubation effects. Seifert et al. (1994) expanded upon this idea, arguing that “failure indices” may become attached to initially unsuccessful retrieval attempts, and chance encounters with related stimuli are more likely to trigger insights or *aha!* experiences. This “opportunistic assimilation” theory, based on unconscious activation of concepts, has been offered as a theory of incubation effects.

Another feature of creative thinking, *intuition*, has been treated as below-threshold activation by Bowers and colleagues (Bowers et al., 1990, 1995). In contrast with insight, an idea one becomes consciously aware of, intuition is treated as a “hunch” by Bowers and colleagues. According to this view, intuitive guiding, rising from unconscious activation, can lead one toward more coherent solutions. A similar aspect of implicit cognition has been examined in terms of predicting insight experiences in problem solving (Metcalfe, 1986; Metcalfe & Weibe, 1987). These studies introduced a metacognitive definition of insight experiences, based on repeated subjective reports known as *warmth ratings*. Because intuition is defined as a hunch, or an ineffable “gut feeling,” warmth ratings might be considered estimates of intuition. For noninsight problems, this metacognitive measure shows regular increases over time until a solution is reached. For insight problems, however, increasing warmth ratings did not predict impending solutions; rather, they predicted impending failures. Metcalfe’s research shows that intuition may not be a good predictor of insight.

Implicit memory in the form of involuntary retrieval of recently encountered stimuli has been

linked to fixation and conformity effects in creative invention and problem solving (e.g., Smith, 1995b, 2003). Smith and Tindell (1997) used word fragment completion to induce implicit memory blocks, priming with words orthographically similar to test fragments, but that could not correctly complete the fragments. The resulting implicit memory blocks (e.g., Kinoshita & Towgood, 2001; Leynes, Rass, & Landau, 2008; Logan & Balota, 2003; Lustig & Hasher, 2001) are not eliminated when participants are explicitly warned not to think about read words while they were completing the test fragments. Parallel results have been reported, showing that similar warnings do not mitigate conformity effects in creative idea generation (Smith et al., 1993) or fixation in creative design (Jansson & Smith, 1991).

Creative Operations, Procedures, and Activities

Although no fixed set of cognitive operations or procedures is common for all creative activities, there are, nonetheless, some that often are encountered and considered in the research literature on creative thinking. These operations include the practice of generating lists of ideas (also known as idea generation, or ideation) and combining ideas.

Combination

One of the challenges people confront as a result of relying heavily on existing knowledge in performing creative tasks is generating something new, something that goes beyond what is already known. Indeed, this requirement also represents a challenge for researchers attempting to explain creativity. How are we to account for the appearance of something new if all people have to work with is their existing knowledge? One process that can help with both challenges is *conceptual combination*, in which people merge together two concepts that were previously

Change from A to B in 3 moves.

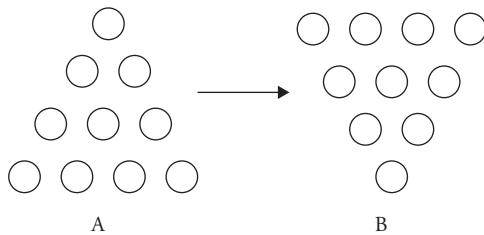


Fig. 23.7 An insight problem, one whose solution tends to be realized suddenly (e.g., Metcalfe, 1986).

completely separate or are otherwise discrepant or dissimilar (e.g., Medin & Shoben, 1988; Shoben & Gagné, 1997; Sifonis & Ward, 1998; Wisniewski, 1997). One of the reasons conceptual combination can be effective in provoking or explaining the origins of novelty is that it can result in the emergence of properties that are not strongly associated with either of the parent concepts in the combination. The creative potential inherent in those *emergent properties* can then serve as the basis for a new ideas or products. A classic example of emergence is seen in the interpretation of the somewhat unusual combination "Harvard-educated carpenter," which might be perceived as being nonmaterialistic, a property not necessarily strongly associated with Harvard-educated people or carpenters in general (e.g., Kunda, Miller, & Claire, 1990). Research is consistent with the idea that the more discrepant the separate concepts are, the more likely it is that emergent properties will be observed (e.g., Estes & Ward, 2002; Hampton, 1987; Wilkenfeld & Ward, 2001), possibly because discrepancy forces people to attempt to resolve the contradiction between the component terms (Hampton, 1997; Kunda et al., 1990).

Ideation

Divergent thinking, the search for many varied and imaginative possible problem solutions, has been contrasted with *convergent thinking*, a type of problem solving or reasoning in which cognitive operations are intended to converge upon the single correct answer to a problem. In divergent thinking tasks, people begin with an ill-defined or open-ended problem, such as finding alternative uses for a common object (e.g., Finke et al., 1992). Performing a divergent thinking task resembles the task of listing members of categories; in both cases, the subject brings all of the required knowledge to the task. There is a tendency to think of these two activities as fundamentally different. Whereas responses in category

generation tasks are often seen as passively residing in a memory repository, ideas for an idea generation task seem to be created on the spot. However, both of these assumptions are flawed. Most categories (e.g., *birds*) have a graded structure, with some members perceived as better members (e.g., *robin*) and others as poorer (e.g., *phoenix*), and ad hoc categories (e.g., *heavy things*), in particular, clearly demand on-the-spot imagination (e.g., *a sumo wrestler's lunchbox*; see Barsalou, 1982, 1985). Divergent thinking, like category generation, requires both retrieval of material from memory and imagination.

When done as a collaborative group, ideation is usually referred to as *brainstorming* (e.g., Osborn, 1957). There are other collaborative methods intended to enhance ideation, but generating and combining ideas is the basic goal of such activities.

Creative Cognition Theories

It is not easy to demarcate theories of creative cognition from theories of cognition. For example, Gentner's structural mapping theory (e.g., Gentner, 1987; Gentner & Markman, 1994; see Holyoak, Chapter 13) deals with analogical reasoning and is relevant to important issues in creative cognition, yet we would not call it a creative cognition theory; rather, it is a theory of cognition that we use to describe aspects of creative thinking. In comparison, Ward's (1994) structured imagination theory was specifically designed to address certain aspects of creative thinking, and it is couched in terms of cognitive theory; therefore, it qualifies as a theory of creative cognition. Here we briefly review theories that were intended to capture important aspects of creative cognition.

Remote Association

The theory of *remote association* (e.g., Mednick, 1962; Mednick, Mednick, & Mednick, 1964), inspired by studies of associative responses given by schizophrenic patients (e.g., Mednick, 1962), took a narrow view of creative thinking as a single associative process. The idea is that the process of accessing or retrieving a remote associate of a stimulus, rather than a common or prepotent response, is a cognitive event at the heart of all truly creative thinking. This process, accessing a remote association, should be useful in the tricky sort of creative problem solving in which the prepotent ("sucker") initial response implicitly includes unnoticed assumptions that prevent successful problem solving. The ability to access a more remote association allows the creative problem solver to go beyond the initial fixation

caused by the prepotent response. Furthermore, remote access allows a person to combine more distantly related ideas or concepts; such unusual combinations might produce novel emergent properties, a hallmark of many creative ideas.

Martindale (1981, 1989, 1995) elaborated and extended the theory of remote association, adding mechanisms such as attentional defocusing (or narrowing) that can impact the gradient of associative hierarchies. A steep associative hierarchical gradient (where prepotent responses are far more likely than other associates to be accessed) can shift to a flatter gradient, in which remote associates are more accessible (Martindale, 1995). Flattening of this associative gradient via defocused attention, according to Martindale, should be directly related to sympathetic arousal levels, because lower arousal acts to increase the randomicity (temperature) function in Martindale's connectionist model, facilitating escape from local minima.

Darwinian Model

A Darwinian model of creativity is one that embraces the notion of random variation and selection at the level of ideas (e.g., Simonton, 1999; see Simonton, Chapter 25). That is, one must continue to produce variations of ideas until, by chance, an idea turns out to be creative and valuable. Key to this theory is the idea that creativity is a stochastic process, rather than a deterministic one. That is, one cannot produce creative products by virtue of a preplanned step-by-step process. Rather, creativity is seen as a lucky accident. Therefore, increasing the chances of a providential accident can be accomplished by increasing the number of ideas that are generated and considered in the course of creative ideation.

This Darwinian theory of creative ideation does not really specify the cognitive mechanisms that determine variations in the quantity of ideas that can be generated and, in fact, does not depend upon the empirical validity of any cognitive mechanisms, except for those that might contribute to the quantity of ideas that are produced. Indeed, the quantity of ideas, often measured as fluency, is one of the most common measures of creative productivity. Most experimental studies of brainstorming, for example, have relied upon measures of the quantity of ideas produced in various conditions. Shah et al. (2003) referred to quantity as a *process* measure, rather than an *outcome* measure of creativity. That is, more ideas might increase the chances that a highly creative

idea will be produced (the Darwinian principle), but quantity is not a necessary outcome if someone only thinks of that single most creative idea.

Opportunistic Assimilation

The "prepared mind" theory, based on Pasteur's notion that "Chance favors the prepared mind," is another cognitive model of aspects of creative thinking. Encompassing the principle of the prepared mind is the *opportunistic assimilation* theory of Seifert and colleagues (e.g., Patalano & Seifert, 1997; Seifert et al., 1994; Seifert & Patalano, 2001). This theory explains insight experiences with a combination of several cognitive mechanisms, including (unconscious) spreading activation, failure indices encoded with unsolved problems, and serendipitous environmental triggers. The theory states that recognizing opportunities to fulfill pending goals is improved by predictively encoding one's goal, which allows people to take advantage of unforeseen opportunities to achieve those goals. When attempts to solve a problem result in initial failures, these episodes can be encoded in memory as associations between pending goals and stimulus features that might potentially solve the failed problem. Unsolved problems that are encoded in association with these "failure indices" may be remembered when stimuli with features related to those indices are serendipitously encountered. Furthermore, unconscious semantic activation of failed problems primes the encoded representations of unsolved problems, making them more likely to be remembered when such chance encounters occur (e.g., Yaniv & Meyer, 1987).

According to the opportunistic assimilation theory, this combination of encoded failure indices and unconscious spreading activation describes the prepared mind, because it prepares the problem solver to be able to use relevant cues without requiring the examination of every object in the environment to consider its relevance to the unsolved problem. Encoding more abstract versions of failure indices, according to this view, is more advantageous than specific or concrete encodings because abstract problem needs might be satisfied by an entire class of objects, rather than requiring a specific object. This opportunistic assimilation was designed to explain incubation effects in creative problem solving: When one initially fails to solve a problem, and the problem is put aside, a serendipitously encountered object in one's environment can provide a relevant clue, triggering an insight.

Idea Roadmaps

Smith's roadmaps theory (Smith, 1995a) provides a theoretical basis for a generative search of one's knowledge, a search for information that could potentially address one's goals and problem solutions. The theory does not assume that goals or problem solutions are necessarily stored in their final form, but rather that goals are achieved via an incremental construction process that can combine preexisting knowledge with new ideas constructed from combinations of existing knowledge. The theory can be applied to creative problem solving, design, or other similar constructive activities.

The multidimensional space in which ideas are constructed, according to an extension of this theory (Smith, 1995b), is determined and bounded by the *plan* adopted to guide thinking. This type of plan includes a set of operations for manipulating the content of ideas that are under construction; these operations, combined with the content that serves as a problem's initial representation, determine an idea space that can be constructed and navigated as the idea is developed. If this idea space is arranged hierarchically, with each location defined in terms of the idea's specified content, then this idea roadmap can be oriented with the least specific content at the top of the roadmap and the most specified ideas at the bottom of the hierarchy. Typical problem solving, according to this theory, tends to begin at or near the top of a roadmap and proceed incrementally toward the bottom of the roadmap, where specific solutions are represented. During creative idea construction, each successive representation of an idea on a roadmap tends to be a precursor for a more specific representation, so there may be a tendency to include each represented set of elements automatically on subsequent problem-solving steps. Thus, representations on the roadmap of ideas tend to encode implicit knowledge, memories, or assumptions about problem solutions.

Fixation, according to this theory, corresponds to taking a dead-end branch in the path that is constructed in the course of creative problem solving. If a problem solution or a creative insight is not located on a projected idea map, and we assume only downward movement within the idea space, this provides a way to view fixation in creative thinking. "Upward" movement on an idea roadmap involves removing elements from idea representations, which can be especially difficult for elements that have become implicitly included components of a representation. Incubation, in this theory, allows one

to escape from dead-end fixated paths. If a break or context shift enables a new idea representation that does not include fixating elements, then fixation may be resolved—an incubation effect, due to restructuring of the problem.

This roadmaps theory blends well with Ward's ideas about structured imagination and his path-of-least-resistance theory, as described earlier. That is, a hierarchically structured conceptual map forms the basis of imagination in both cases. In the structured imagination view, beginning ideation with a representation that is too concretely specified can limit the breadth of one's search for ideas. In the roadmaps view, fixation and conformity due to recent experience can likewise limit the breadth of idea generation.

Metrics of Creativity

How can we recognize creativity when we see it? How can we measure creativity? The same basic questions have been asked about other cognitive domains, such as memory. Cued recall and word stem completion are two different measures commonly used to measure memory; which is a better measure of the construct? Most cognitive psychologists understand the absurdity of this question, because the answer begins with "It depends on what you want to measure." There is no "single" or "best" measure of memory, because memory involves many complex interacting systems. The same can be said about creativity. What is the best measure of creativity? It depends on what aspect of creative thinking you want to measure. Studies of creative thinking often focus on specific aspects of creative cognition, such as creative problem solving, idea generation, conceptual combination, or visualization.

Although many studies have looked at performance on insight problems to measure creative problem solving (e.g., Maier, 1931; Metcalfe, 1986; Smith & Blankenship, 1989, 1991), studies of creative thinking must go beyond puzzles to connect cognition with creative products. In these cases, the most typical metrics used to measure creative output are the *quantity*, *quality*, *variety*, and *novelty* of responses or products. For example, in a divergent thinking task in which one is asked to list the alternative uses of empty 2-liter soda bottles, the number of ideas listed is the quantity, the number of categories of ideas (e.g., construction, weapon, things that float) is the variety, and the statistical infrequency of each idea, as measured by a norm, is the novelty. Quality is usually subjectively judged. Shah et al. (2000) distinguished *process metrics* from

outcome metrics of creative ideation, stating that the novelty and quality are important for judging the outcome or end product of creative work, whereas quantity and variety are important only for judging aspects of the creative process. These four metrics, and variants of them, have proven useful in different ideation contexts, such as divergent thinking tasks (e.g., Guilford, 1967), brainstorming (e.g., Kohn & Smith, 2010), and engineering design (e.g., Vargas-Hernandez, Shah, & Smith, 2010). Nonetheless, there are important limitations of these creativity metrics. One limitation has to do with domain differences and idiosyncrasies (e.g., engineering design may have many important creative needs that differ from those of music composition, business, and science).

Some relatively new metrics of creative cognition have been identified, including measures of conformity, emergence, and abstraction. Smith et al. (1993) defined *conformity* as using features of examples one has seen in one's own creative ideas. They and others (e.g., Dahl & Moreau, 2002; Jansson & Smith, 1991; Vargas-Hernandez et al., 2010) have shown that conformity may not be overcome by instructions to avoid using features of examples. Kohn and Smith (2010) showed that brainstorming participants tend to conform to the ideas of others in their brainstorming group. Although the term "conformity" typically is used pejoratively, it is clearly the case that conforming to useful examples in education and training is quite useful and important. When examples block or unnecessarily limit creative thinking, however, conformity is a problem. *Emergence* has been measured as a property seen in creative thinking in which combinations of ideas can produce concepts with qualities that are not seen in the component concepts (e.g., Estes & Ward, 2002; Wilkenfeld & Ward, 2001), and it has been used as a measure of creative cognition (e.g., Kerne, Smith, Koh, Choi, & Graeber, 2008a). Another metric of creative thinking is *abstraction*, that is, progressing from relatively concrete representations to more general levels, which potentially enables a broader variety of ideas. Abstraction can circumvent fixation and overly structured imagination (e.g., Ward et al., 2004), and it may promote access to remote associations and analogies.

Impediments and Aids to Creative Thinking

There are several known impediments and aids to creative thinking that relate directly to patterns

of cognition. Here, we briefly review some of these impediments and aids.

Impediments

INADEQUATE KNOWLEDGE

Perhaps the most common impediment to creative thinking is a lack of knowledge and resources (e.g., Weisberg, 1992, 2006). Both everyday creativity and extraordinary creativity (see Simonton, Chapter 25) require prior knowledge and experience. A myth of creativity is that one needs no expertise, and that an active imagination is all one needs to create great things; such notions are nonsense. A creative imagination is necessarily structured according to one's prior knowledge (e.g., Ward, 1994), and the value of serendipitous discoveries can only be noticed or realized by those with some level of expertise in a domain.

IMPLICIT ASSUMPTIONS

Although knowledge and experience are necessary for creative thinking, they are not sufficient in many identifiable cases. Inappropriate ideas or approaches to a problem can impede creative thinking via fixation (e.g., Smith & Blankenship, 1989, 1991; Wiley, 1998) and the implicit use of inappropriate assumptions (e.g., Smith, 1994). Findings that show fixation and implicit assumptions (e.g., Luchins & Luchins, 1959) can block creative cognition and demonstrate that prior experience can, at times, have negative consequences (see Fig 23.8). What is insidious about the implicit assumptions that can cause fixation is that such impediments are hidden from one's conscious mind, making them difficult to detect and remedy.

COGNITIVE ILLUSIONS

Cognitive illusions, such as false memories (e.g., Roediger & McDermott, 1995), memory misattributions (e.g., Jacoby, 1991), or misapplied availability heuristics (e.g., Tversky & Kahneman, 1974), cause predictable, systematic cognitive errors. One cognitive illusion that can impede creative thinking is *hindsight bias*. The sense that retrospective views of ideas make them seem more obvious can impede certain aspects of creativity (e.g., Fischhoff, 1975). A good example of hindsight bias can be seen in the patent process. A U.S. patent can be granted only for inventions that are "nonobvious" to practitioners of a profession. This nonobviousness criterion may lead to denial of patents for creative ideas that, in retrospect, might seem obvious to patent examiners (e.g., Seifert, 2008; Smith, 2008).

Problem #	Jar A	Jar B	Jar C	Goal Amount	Solution
1	64	100	6	24	B - A - 2C
2	21	73	5	42	
3	6	28	7	8	
4	35	94	21	17	
5	22	49	12	3	
6	7	56	8	33	
7	15	39	3	18	
8	28	76	3	25	

Fig. 23.8 Luchins' water jar problem: Using only the jars A, B, and C as measuring devices, measure out the desired amount for each problem. The mental set B – A – 2C, learned early in the problem sequence, may become mindlessly reapplied on each successive problem. This fixation effect, that is, blind use of a mental set, leads to use of a complicated solution (B – A – 2C) where a simple one should be apparent on Problem 7 (A + C), and it leads to the use of an incorrect solution (B – A – 2C) on Problem 8, even though a simple correct solution (A – C) should be readily apparent.

PREMATURE CONCEPTUALIZATION

One may begin creative work on an idea with a relatively specific or concrete representation, or one may reach such a representation early on in the process. Taking the path of least resistance, one may thereafter implicitly include all of those prematurely specified features in subsequently developed versions of a creative idea, never considering ideas that lack those features.

Aids

Creativity can be aided in many ways, such as gaining knowledge and expertise, using analogies, combining ideas, thinking abstractly, redefining problems that are fixated, and noticing ways in which a new idea could have important implications.

COMBINATION

Ever since people have tried to encourage creative thinking, the notion has flourished that combinations of existing ideas can form the basis of creative ideas (e.g., Osborn, 1957). As more interpretations of a conceptual combination are generated, the interpretations increase in their normative originality, and they tend to be based more on idiosyncratic knowledge of the components of the combination. One of the main goals in creative conceptual combination is to discover new and potentially useful emergent properties: that is, properties not commonly seen in the component concepts, but that emerge only in combinations (e.g., Estes & Ward, 2002).

ANALOGY

Because analogies link conceptual domains that are similar, they can provide vehicles for constructing creative solutions that cannot be found within the domain of a problem (e.g., Gick & Holyoak, 1980; see Holyoak, Chapter 13). Typically, the most useful analogies will be those that are based on the similarity of deeper meaningful levels of problem topics and solution vehicles, as opposed to similarity of superficial features of topic and vehicle domains. Local analogies are those that are conceptually closer to a problem topic domain than are remote analogies. Like remote associations, remote analogies appear to be more useful than local analogies for creative design (e.g., Christensen & Schunn, 2007), whereas Dunbar (e.g., 1997) found the reverse for a group of scientists engaged in scientific problem solving (also see Dunbar & Klahr, Chapter 35).

ABSTRACTION

Because imagination is structured according to underlying concepts in memory (e.g., Ward, 1994), beginning creative idea development at a convenient basic level and then taking the path of least resistance can lead to premature conceptualization. Abstract thinking, particularly early on in the ideation process, can lead one to explore a greater range of ideas, effectively expanding the conceptual space under consideration.

NOTICING

Stumbling across a key clue is not sufficient; one must *notice* the implications of that clue and realize its implications in terms of creative products or unsolved problems. No mind can be prepared to notice such clues without sufficient expert knowledge. Noticing may be facilitated by frustration from initial failures, which potentiates unsolved problems, helping them come to mind when relevant clues appear (e.g., Patalano & Seifert, 1997; Seifert et al., 1994).

KNOWLEDGE

Ideally, an expert always knows the correct approach for any given problem. Even experts, however, are susceptible to fixation and conformity effects in inventive design tasks (e.g., Jansson & Smith, 1991; Linsey et al., 2009). Redefinition of a problem or goal can sometimes provide a key insight, a restructuring of the problem, particularly when problems cannot be solved using expert

heuristics. One way to aid redefinition is via perspective shifts: thinking about problems in new contexts, ones that may not be associated with fixated approaches. Another aid to redefinition is analogy (e.g., Linsey, Wood, & Markman, 2008).

SUPPORT TECHNOLOGIES

Although there are many physical tools and technologies that aid the cognitive structures and operations that give rise to creativity, sketching is one of the oldest and most universal of these external aids (see Hegarty & Stull, Chapter 31). Sketching provides a particularly good medium for expressing and sharing spaces, schemas, and mental models (Tversky & Hard, 2001), freeing up cognitive resources, and extending our visualization abilities in ways that are particularly conducive to creative cognition (Tversky, 2001, 2005). Because sketches can be reexamined and reconsidered, they can support restructuring and discovery of unanticipated relations (Goel, 1995; Goldschmidt, 1994; Suwa & Tversky, 2001; Tversky & Suwa, 2009).

Conclusions

Creative thinking involves all of the cognitive systems. There is no single creative process; rather, there are many different types of creative ideas, and many ways in which ideas can be generated, constructed, and developed. Many of these pathways to creativity can be described and explained in terms of the cognition that underlies the creation of ideas. As in most other areas of cognitive psychology, the creative cognition approach has used empirical science to study the nature and functioning of cognitive processes and structures that underlie creative thinking. Cognitive operations that often are involved in creative cognition include divergent idea generation, the formation of conceptual combinations, retrieval and mapping of analogies, abstraction, visualization, and conceptual restructuring. Common impediments to creative cognition include inadequate prior knowledge, implicit assumptions, cognitive illusions, and premature conceptualization in the development of creative ideas. Methods and tools that can support creative cognition must be based on the cognitive operations and structures that give rise to creative ideas.

Future Directions

Cognitive Neuroscience

The role of the brain and nervous system in creative cognition is beginning to be explored by some

cognitive neuroscientists, and there are many ways in which neuroscientific mechanisms might contribute to creative thinking. Martindale, for example, linked access to remote associates with the effects that low physiological arousal have on neural communication (e.g., Martindale, 1981, 1995). Another excellent example is the work of Jung-Beeman, Kounios, and others who have developed a theory of insight and comprehension based on functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) studies of participants solving insight problems (e.g., Jung-Beeman, 2005; Jung-Beeman et al., 2004; Kounios et al., 2006; Kounios et al., 2008). These studies have found, for example, activity in the right anterior superior temporal gyrus linked with awareness of a solution to an insight problem (e.g., Jung-Beeman et al., 2004; see van Steenburgh et al., Chapter 24). Future research on the neuroscience of creative cognition must blend neuroscience with cognitive mechanisms that are well known for their link to creative thinking, as the aforementioned studies have already begun to do.

Collaborative Cognition

Much of the cognition involved in real-world creativity is done collaboratively, using multiple types of media. Many studies have documented brainstorming deficits when participants are in groups: More ideas are produced when the individuals brainstorm individually (e.g., Diehl & Stroebe, 1987, 1991; Kohn & Smith, 2010; Nijstad et al., 2003; Nijstad & Stroebe, 2006). Cognitive mechanisms such as fixation and conformity to the ideas of others have been shown to limit the number and the novelty of ideas of brainstorming participants (Kohn & Smith, 2010). Parallel findings of collaborative inhibition in memory recall show that more is recalled when participants recall events individually (e.g., Weldon & Bellinger, 1997). Future research must pursue questions about the cognition that occurs in groups and how that cognition affects creativity. Particularly relevant to questions about collaborative creativity is the increasing role of the Internet in customer-driven design, and virtual teams (e.g., Maher, 2010; Ward, Guerdat, & Roskos, 2010).

Digital Tools

Digital tools that can aid creativity can be considered extensions of the types of human cognition that participate in creative cognition (Smith, Kerne, Koh, & Shah, 2009) (see Fig. 23.9). These digital tools can promote, for example, analogical

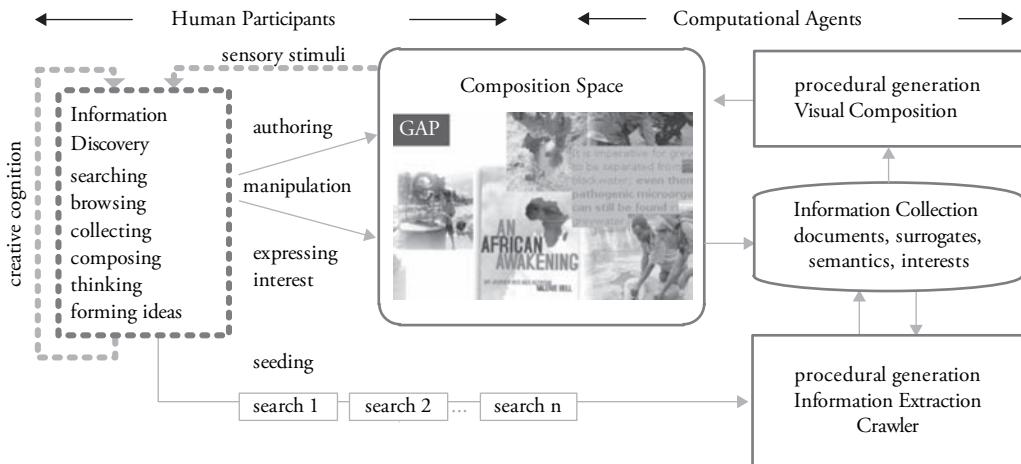


Fig. 23.9 Computational agents support human participants in the combinFormation digital tool for information discovery, which involves browsing through collections returned by search engines and forming navigable compositions of relevant results (Kerne, Koh, Smith, Webb, & Dworaczyk, 2008b). Problems are iteratively reformulated, representations are manipulated, and solutions constructed, often involving integration of multiple information resources (Kerne & Smith, 2004). See color figure.

reasoning (Markman, Wood, Linsey, Murphy, & Laux, 2009), conceptual combination of remote associates (Kerne et al., 2008a), or abstract thinking (Ward, 2009). Future research in creative cognition should examine this fundamental link between cognition and digital tools designed to support and enhance creativity.

Creative Expertise

Do some people gain creative expertise; that is, do they become expert at recognizing blocks, do they know good methods for escaping blocks, redefining problems, combining and transforming concepts, drawing deeply appropriate but remote analogies, and so on? Is this type of creative expertise something that can be effectively learned and trained? Whether such putative expertise actually enhances creativity is an important question for future research in creative cognition.

References

- Barsalou, L. W. (1982). Context-independent and context-dependent information in concepts. *Memory and Cognition*, 10, 82–93.
- Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11, 629–654.
- Beghetto, R., & Kaufman, J. (2007). Toward a broader conception of creativity: A case for ‘mini-c’ creativity. *Psychology of Aesthetics, Creativity, and the Arts*, 1, 73–79.
- Bower, G. H., Clark, M., Winzenz, D., & Lesgold, A. (1969). Hierarchical retrieval schemes in recall of categorized word lists. *Journal of Verbal Learning and Verbal Behavior*, 8, 323–343.
- Bowers, K. S., Farvolden, P., & Mermigis, L. (1995). Intuitive antecedents of insight. In S. M. Smith, T. B. Ward, & R. A. Finke (Eds.), *The creative cognition approach*. (pp. 27–51). Cambridge, MA: MIT Press.
- Bredart, S., Ward, T. B., & Marczewski, P. (1998). Structured imagination of novel creatures’ faces. *American Journal of Psychology*, 111, 607–725.
- Brown, A. S., & Hall, L. A. (1979). Part-list cuing inhibition in semantic memory structures. *American Journal of Psychology*, 92, 351–362.
- Choi, H., & Smith, S. M. (2005). Incubation and the resolution of Tip-of-the-tongue states. *The Journal of General Psychology*, 132(4), 365–376.
- Christensen, B. T., & Schunn, C. D. (2007). The relationship between analogical distance to analogical function and pre-inventive structure: The case of engineering design. *Memory and Cognition*, 35, 29–38.
- Chrysikou, E. G., & Weisberg, R. W. (2005). Following the wrong footsteps: Fixation effects of pictorial examples in a design problem-solving task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5), 1134–1148.
- Connor, L. T., Balota, D. A., & Neely, J. H. (1992). On the relation between feeling of knowing and lexical decision: Persistent subthreshold or topic familiarity? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(3), 544–554.
- Dahl, D. W., & Moreau, P. (2002). The influence and value of analogical thinking during new product ideation. *Journal of Marketing Research*, 39, 47–60.
- Diehl, M., & Stroebe, W. (1987). Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of Personality and Social Psychology*, 53(3), 497–509.
- Diehl, M., & Stroebe, W. (1991). Productivity loss in idea-generating groups: Tracking down the blocking effect. *Journal of Personality and Social Psychology*, 61(3), 392–403.
- Dominowski, R. L. (1981). Comment on “An examination of the alleged role of ‘fixation’ in the solution of ‘insight’ problems.” *Journal of Experimental Psychology: General*, 110, 199–203.

- Dominowski, R. L., & Jenrick, R. (1972). Effects of hints and interpolated activity on solution of an insight problem. *Psychonomic Science*, 26(6), 335–338.
- Dunbar, K. (1997). How scientists think: On-line creativity and conceptual change in science. In T. B. Ward, S. M. Smith, & J. Vaid (Eds.), *Creative thought: An investigation of conceptual structures and processes* (pp. 461–493). Washington, DC: American Psychological Association.
- Duncker, K. (1945). On problem solving. *Psychological Monographs*, 58, (5, Whole No. 270).
- Ellen, P. (1982). Direction, past experience, and hints in creative problem solving: Reply to Weisberg and Alba. *Journal of Experimental Psychology: General*, 111, 316–313.
- Erdelyi, M. H., & Becker, J. (1974). Hypernesia for pictures: Incremental memory for pictures but not words in multiple recall trials. *Cognitive Psychology*, 6(1), 159–171.
- Estes, Z., & Ward, T. B. (2002). The emergence of novel attributes in concept modification. *Creativity Research Journal*, 14, 149–156.
- Finke, R., Ward, T., & Smith, S. M. (1992). *Creative cognition*. Cambridge, MA: MIT Press.
- Fischhoff, B. (1975). Hindsight foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, 1, 288–299.
- Gelman, S. A. (1988). The development of induction within natural kind and artifact categories. *Cognitive Psychology*, 20, 65–95.
- Gentner, D. (1987). Structure mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155–170.
- Gentner, D., & Markman, A. B. (1994). Structural alignment in comparison: No difference without similarity. *Psychological Science*, 5, 152–158.
- Gick, M. L. & Holyoak, K. L. (1980). Analogical problem solving. *Cognitive Psychology*, 15, 306–355.
- Glucksberg, S., & Weisberg, R. W. (1966). Verbal behavior and problem solving: Some effects of labelling in a functional fixedness problem. *Journal of Experimental Psychology*, 71, 659–664.
- Goel, V. (1995). *Sketches of thought*. Cambridge, MA: MIT Press.
- Goldschmidt, G. (1994). On visual design thinking: The vis kids of architecture. *Design Studies*, 15(2), 158–174.
- Guilford, J. P. (1967). *The nature of human intelligence*. New York: McGraw-Hill.
- Hampton, J. A. (1987). Inheritance of attributes in natural concept conjunctions. *Memory and Cognition*, 15, 55–71.
- Hampton, J. A. (1997). Emergent attributes in combined concepts. In T. B. Ward, S. M. Smith, & J. Vaid (Eds.), *Creative thought: An investigation of conceptual structures and processes* (pp. 83–110). Washington, DC: APA Books.
- Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin and Review*, 7, 569–592.
- Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, 30, 513–541.
- Jansson, D. G., & Smith, S. M. (1991). Design fixation. *Design Studies*, 12(1), 3–11.
- Jones, G. V. (1989). Back to Woodworth: Role of interlopers in the tip of the tongue phenomenon. *Memory and Cognition*, 17, 69–76.
- Jones, G. V., & Langford, S. (1987). Phonological blocking in the tip of the tongue state. *Cognition*, 26, 115–122.
- Jung-Beeman, M. (2005). Bilateral brain processes for comprehending natural language. *Trends in Cognitive Sciences*, 9, 512–518.
- Jung-Beeman, M., Bowden, E. M., Haberman, J., Frymiare, J. L., Arambel-Liu, S., & Greenblatt, R. (2004). Neural activity when people solve verbal problems with insight. *PLoS Biology*, 2, 500–510.
- Kaufman, J. C., & Beghetto, R. A. (2009). Beyond big and little: The Four C model of creativity. *Review of General Psychology*, 13, 1–12.
- Kerne, A., Smith, S. M., Koh, E., Choi, H., & Graeber, R. (2008a). An experimental method for measuring the emergence of new ideas in information discovery. *International Journal of Human-Computer Interaction*, 24(5), 460–477.
- Kerne, A., Koh, E., Smith, S. M., Webb, A., & Dworaczyk, B. (2008b). combinFormation: Mixed-initiative composition of image and text surrogates promotes information discovery. *ACM Transactions on Information Systems (TOIS)*, 27(1), 1–45.
- Kinoshita, S., & Towgood, K. (2001). Effects of dividing attention on the memory-block effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 889–895.
- Kohn, N. W., & Smith, S. M. (2009). Partly versus completely out of your mind: Effects of incubation and distraction on resolving fixation. *Journal of Creative Behavior*, 43(2), 102–118.
- Kohn, N. W., & Smith, S. M. (2010). Collaborative fixation: Effects of others' ideas on brainstorming. *Applied Cognitive Psychology*, 24, 1–22.
- Kornell, N., & Metcalfe, J. (2006). "Blockers" do not block recall during tip-of-the-tongue states. *Metacognition Learning*, 1, 248–261.
- Kounios, J., Fleck, J. I., Green, D. L., Payne, L., Stevenson, J. L., Bowden, M., & Jung-Beeman, M. (2008). The origins of insight in resting state brain activity. *Neuropsychologia*, 46, 281–291.
- Kounios, J., Frymiare, J. L., Bowden, E. M., Fleck, J.I., Subramaniam, K., Parrish, T. B., & Jung-Beeman, M. J. (2006). The prepared mind: Neural activity prior to problem presentation predicts subsequent solution by sudden insight. *Psychological Science*, 17, 882–890.
- Kunda, Z., Miller, D. T., & Claire, T. (1990). Combining social concepts: The role of causal reasoning. *Cognitive Science*, 14, 551–577.
- Landau, J. D., & Lehr, D. P. (2004). Conformity to experimenter-provided examples: Will people use an unusual feature? *Journal of Creative Behavior*, 38(3), 180–191.
- Leynes, P. A., Rass, O., & Landau, J. D. (2008). Eliminating the memory blocking effect. *Memory*, 16(8), 852–872.
- Linsey, J., Tseng, I., Fu, K., Cagan, J., & Wood, K. (2009, August). Reducing and perceiving design fixation: Initial results from a NSF-sponsored workshop. Paper presented at the International Conference on Engineering Design, Stanford, CA.
- Linsey, J., Wood, K., & Markman, A. (2008, August). Increasing innovation: Presentation and evaluation of the word tree design-by-analogy method. Paper presented at the ASME IDETC Design Theory and Methodology Conference, New York, NY.
- Logan, J. M., & Balota, D. A. (2003). Conscious and unconscious lexical retrieval blocking in younger and older adults. *Psychology and Aging*, 18, 537–550.
- Luchins, A., & Luchins, E. (1959). *Rigidity of behavior: A variational approach to the effect of einstellung*. Eugene: University of Oregon Books.

- Lustig, C., & Hasher, L. (2001). Implicit memory is vulnerable to proactive interference. *Psychological Science*, 12, 408–412.
- Maher, M. L. (2010). Design creativity research: From the individual to the crowd. In T. Taura & Y. Nagai (Eds.), *Design creativity 2010* (pp. 41–48). London: Springer.
- Maier, N. R. F. (1930). Reasoning in humans: I. On direction. *Comparative Psychology*, 12, 115–143.
- Maier, N. R. F. (1931). Reasoning in humans: II. The solution of a problem and its appearance in consciousness. *Comparative Psychology*, 12, 181–194.
- Markman, A. B., Wood, K. L., Linsey, J. S., Murphy, J. T., & Laux, J. P. (2009). Supporting innovation by promoting analogical reasoning. In A. B. Markman & K. L. Wood (Eds.), *Tools for innovation* (pp. 85–103). New York: Oxford University Press.
- Martindale, C. (1981). *Cognition and consciousness*. Homewood, IL: Dorsey.
- Martindale, C. (1989). Personality, situation, and creativity. In J. A. Glover, R. R. Ronning, & C. R. Reynolds (Eds.) *Handbook of creativity* (pp. 211–228). New York: Plenum.
- Martindale, C. (1995). Creativity and connectionism. In S. M. Smith, T. B. Ward, & R. A. Finke (Eds.), *The creative cognition approach* (pp. 249–268). Cambridge, MA: MIT Press.
- Medin, D. L., & Shoben, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology*, 20, 158–190.
- Mednick, M. T., Mednick, S. A., & Mednick, E. V. (1964). Incubation of creative performance and specific associative priming. *Journal of Abnormal and Social Psychology*, 69, 84–88.
- Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, 69(3), 220–232.
- Metcalf, J., & Weibe, D. (1987). Intuition in insight and noninsight problem solving. *Memory and Cognition*, 15(3), 238–246.
- Metcalf, J. (1986). Premonitions of insight predict impending error. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12, 623–634.
- Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92, 289–316.
- Nijstad, B. A., & Stroebe, W. (2006). How the group affects the mind: A cognitive model of idea generation in groups. *Personality and Social Psychology Review*, 10, 186–213.
- Nijstad, B. A., Stroebe, W., & Lodewijkx, H. F. M. (2003). Production blocking and idea generation: Does blocking interfere with cognitive processes? *Journal of Experimental Social Psychology*, 39(6), 531–548.
- Osborn, A. (1957). *Applied imagination*. New York: Scribner.
- Oserson, D. N., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, 97, 185–200.
- Patalano, A. L., & Seifert, C. M. (1997). Opportunism in planning. *Cognitive Psychology*, 34, 1–36.
- Reason, J., & Lucas, D. (1983). Using cognitive diaries to investigate naturally occurring memory blocks. In J. E. Harris & P. E. Morris (Eds.), *Everyday memory: Actions and absentmindedness* (pp. 53–69). London: Academic Press.
- Reitman, W. (1965). *Cognition and thought*. New York: Wiley.
- Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior*, 14, 665–681.
- Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 803–814.
- Roediger, H. L., III, & Thorpe, L. A. (1978). The role of recall time in producing hypermnesia. *Memory and Cognition*, 6(3), 296–305.
- Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, 4, 328–350.
- Rosch, E. H. (1975). Cognitive reference points. *Cognitive Psychology*, 7, 532–547.
- Rosch, E. H., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8, 382–439.
- Schooler, J. W., & Melcher, J. (1995). The ineffability of insight. In S. M. Smith, T. B. Ward, & R. A. Finke (Eds.), *The creative cognition approach* (pp. 97–134). Cambridge, MA: MIT Press.
- Seifert, C. M. (2008). Now why didn't I think of that? The cognitive processes that create the obvious. *Lewis & Clark Law Review*, 12(2), 489–507.
- Seifert, C. M., & Patalano, A. L. (2001). Opportunism in memory: Preparing for chance encounters. *Current Directions in Psychological Science*, 10(6), 198–201.
- Seifert, C. M., Meyer, D. E., Davidson, N., Patalano, A. L., & Yaniv, I. (1994). Demystification of cognitive insight: Opportunistic assimilation and the prepared-mind perspective. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight* (pp. 65–124). Cambridge, MA: MIT Press.
- Shah, J. J., Kulkarni, S. V., & Vargas-Hernandez, N. (2000). Evaluation of idea generation methods for conceptual design: Effectiveness metrics and design of experiments. *Journal of Mechanical Design*, 122, 377–384.
- Shah, J. J., Smith, S. M., & Vargas-Hernandez, N. (2003). Metrics for measuring ideation effectiveness. *Design Studies*, 24, 111–134.
- Shoben, E. J., & Gagne, C. L. (1997). Thematic relations and the creation of combined concepts. In T. B. Ward, S. M. Smith, & J. Vaid (Eds.), *Creative thought: An investigation of conceptual structures and processes* (pp. 31–50). Washington, DC: APA Books.
- Sifonis, C. M., & Ward, T. B. (1998). Structural alignment in relational interpretations of conceptual combinations. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 968–973). Hillsdale, NJ: Erlbaum.
- Simonton, D. K. (1999). *Origins of genius: Darwinian perspectives on creativity*. Oxford, England: Oxford University Press.
- Smith, S. M. (1994). Getting into and out of mental ruts: A theory of fixation, incubation, and insight. In R. J. Sternberg & J. Davidson (Eds.), *The nature of insight* (pp. 121–149). Cambridge, MA: MIT Press.
- Smith, S. M. (1995a). Creative cognition: Demystifying creativity. In C. N. Hedley, P. Antonacci, & M. Rabinowitz (Eds.), *The mind at work in the classroom: Literacy and thinking* (pp. 31–46). Hillsdale, NJ: Erlbaum.
- Smith, S. M. (1995b). Fixation, incubation, and insight in memory, problem solving, and creativity. In S. M. Smith, T. B. Ward, & R. A. Finke (Eds.), *The creative cognition approach* (pp. 135–155). Cambridge: MIT Press.
- Smith, S. M. (2003). The constraining effects of initial ideas. In P. Paulus & B. Nijstad (Eds.), *Group creativity: Innovation through collaboration* (pp. 15–31). New York: Oxford University Press.
- Smith, S. M. (2008). Invisible assumptions and the unintentional use of knowledge and experiences in creative cognition. *Lewis and Clark Law Review*, 12(2), 101–116.

- Smith, S. M., & Blankenship, S. E. (1989). Incubation effects. *Bulletin of the Psychonomic Society*, 27(4), 311–314.
- Smith, S. M., & Blankenship, S. E. (1991). Incubation and the persistence of fixation in problem solving. *American Journal of Psychology*, 104, 61–87.
- Smith, S. M., Kerne, A., Koh, E., & Shah, J. (2009). The development and evaluation of tools for creativity. In A. B. Markman & K. L. Wood (Eds.), *Tools for innovation* (pp. 128–152.) New York: Oxford University Press.
- Smith, S. M., & Tindell, D. R. (1997). Memory blocks in word fragment completion caused by involuntary retrieval of orthographically similar primes. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 23(2), 355–370.
- Smith, S. M., & Vela, E. (1991). Incubated reminiscence effects. *Memory and Cognition*, 19(2), 168–176.
- Smith, S. M., Ward, T. B., & Schumacher, J. S. (1993). Constraining effects of examples in a creative generation task. *Memory and Cognition*, 21, 837–845.
- Storm, B. C., & Angello, G. (2010). Overcoming fixation: Creative problem solving and retrieval-induced forgetting. *Psychological Science*. doi: 10.1177/0956797610379864.
- Suwa, M., & Tversky, B. (2001). Constructive perception in design. In J. S. Gero & M. L. Maher (Eds.), *Computational and cognitive models of creative design* (pp. 227–239). Sydney, Australia: University of Sydney.
- Tversky, B. (2001). Spatial schemas in depictions. In M. Gattis (Ed.), *Spatial schemas and abstract thought* (pp. 79–111). Cambridge, MA: MIT Press.
- Tversky, B. (2005). Functional significance of visuospatial representations. In P. Shah & A. Miyake (Eds.), *Handbook of higher-level visuospatial thinking* (pp. 1–34). Cambridge: Cambridge University Press.
- Tversky, B., & Hard, B. M. (2001). Embodied and disembodied cognition: Spatial perspective-taking. *Cognition*, 110, 124–129.
- Tversky, B., & Suwa, M. (2009). Thinking with sketches. In A. B. Markman & K. L. Wood (Eds.), *Tools for innovation: The science behind the practical methods that drive new ideas* (pp. 75–84). New York: Oxford University Press.
- Vargas-Hernandez, N., Shah, J. J., & Smith, S. M. (2010). Understanding design ideation mechanisms through multilevel aligned empirical studies. *Design Studies*, doi:10.1016/j.destud.2010.04.001.
- Vul, E., & Pashler, H. (2007). Incubation benefits only after people have been misdirected. *Memory and Cognition*, 35(4), 701–710.
- Ward, T. B. (1994). Structured imagination: The role of conceptual structure in exemplar generation. *Cognitive Psychology*, 27, 1–40.
- Ward, T. B. (1995). What's old about new ideas? In S. M. Smith, T. B. Ward, & R. A. Finke (Eds.), *The creative cognition approach* (pp. 157–178). Cambridge, MA: MIT Press.
- Ward, T. B. (2009). ConceptNets for flexible access to knowledge. In A. B. Markman & K. L. Wood (Eds.), *Tools for innovation* (pp. 153–170). New York: Oxford University Press.
- Ward, T. B., Dodds, R. A., Saunders, K. N., & Sifonis, C. M. (2000). Attribute centrality and imaginative thought. *Memory and Cognition*, 28, 1387–1397.
- Ward, T. B., Guerdat, M. G., & Roskos, B. (2010, November). *Avatar visibility as a potential detrimental factor in virtual brainstorming*. Paper presented at the Annual Meeting of the Psychonomic Society, St. Louis, MO.
- Ward, T. B., Patterson, M. J., & Sifonis, C. (2004). The role of specificity and abstraction in creative idea generation. *Creativity Research Journal*, 16, 1–9.
- Ward, T. B., Patterson, M. J., Sifonis, C. M., Dodds, R. A., & Saunders, K. N. (2002). The role of graded structure in imaginative thought. *Memory and Cognition*, 30, 199–216.
- Ward, T. B., & Sifonis, C. M. (1997). Task demands and generative thinking: What changes and what remains the same. *Journal of Creative Behavior*, 31, 18–32.
- Weisberg, R. W. (1992). Metacognition and insight during problem solving: Comment on Metcalfe. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 426–431.
- Weisberg, R. W. (2006). Modes of expertise in creative thinking: Evidence from case studies. In K. A. Ericsson, N. Charness, P. Feltovich, & R. R. Hoffman (Eds.), *Cambridge handbook of expertise and expert performance* (pp. 761–787). Cambridge, England: Cambridge University Press.
- Weldon, M. S., & Bellinger, K. D. (1997). Collective memory: Collaborative and individual processes in remembering. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 1160–1175.
- Wiley, J. (1998) Expertise as mental set: The effects of domain knowledge in creative problem solving. *Memory and Cognition*, 26, 716–730.
- Wilkenfeld, M. J., & Ward, T. B. (2001). Similarity and emergence in conceptual combination. *Journal of Memory and Language*, 45, 21–38.
- Wisniewski, E. J. (1997). Conceptual combination: Possibilities and esthetics. In T. B. Ward, S. M. Smith, & J. Vaid (Eds.), *Creative thought: An investigation of conceptual structures and processes* (pp. 51–81). Washington, DC: APA Books.
- Yaniv, I., & Meyer, D. E. (1987). Activation and metacognition of inaccessible stored information: Potential bases for incubation effects in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 187–205.

J. Jason van Steenburgh, Jessica I. Fleck, Mark Beeman, and John Kounios

Abstract

In the century since the Gestalt psychologists introduced insight as a component process of perception and problem solving, researchers have studied the phenomenological, behavioral, and neural components of insight. Whether and how insight is different from other types of problem solving, such as analysis, has been a topic of considerable interest and some contention. In this chapter we develop a working definition of insight and detail the history of insight research by focusing on questions about the influence of the problem solver's prior knowledge, the origins and significance of representational change, and the roles of impasse and incubation. We also review more recent investigations of the neurological correlates of insight, discuss neurobehavioral states that facilitate or inhibit insightful problem solving, and highlight new methods and techniques that are proving useful in extending our knowledge of insight.

Key Words: knowledge, fixation, impasse, restructuring, Gestalt, special process, hemispheric differences, anterior cingulate cortex, superior temporal gyrus

Insight Defined

Insight occurs when a new interpretation of a situation or the solution to a problem suddenly springs into conscious awareness, seems obviously correct, and is accompanied by a surprising and emotional experience known as the "*Aha*" phenomenon (Kaplan & Simon, 1990). Although the necessity of some of these components of insight has been disputed, most researchers and lay people agree that these are at least typical characteristics of insight. To this basic definition theorists often add the requirement that the problem solver has to restructure or change his or her thinking about some aspect of the problem or the solution in order to achieve insight. Insight is usually contrasted with "analytic" solving in which the solver consciously and deliberately manipulates problem elements to discover the solution. Before discussing the mechanisms of insight, it is important to consider whether it is a unique

process or simply a peculiar manifestation of typical problem-solving mechanisms.

Is Insight Different?

When a chapter on insight appears within a volume that includes a separate chapter on problem solving (see Bassok & Novick, Chapter 21), it is implicitly assumed that insight problem solving is fundamentally different from other types of problem solving. However, the possibility that insightful problem-solving processes share the same mechanisms as analytic processes must be considered. Theorists belonging to the "business-as-usual" camp have argued that the processes by which problems are solved via insight are the same as those used in search solutions and that it is only the affective experience that is different (Atchley, Keeney, & Burgess, 1999; Perkins, 1981; Weisberg 1986, 2006; Weisberg & Alba, 1981). In research examining

verbal protocols, Perkins reported that participants experienced insight characteristics, such as the affectively loaded *Aha* reaction, in conjunction with analytic-type search-based solutions. More recently, researchers have noted that the heuristics traditionally applied during the solution of such search problems (e.g., hill climbing and means-ends analysis, which select moves that appear to make progress toward the solution) can be used to explain the processing displayed by participants when solving with insight (Chronicle, MacGregor, & Ormerod, 2004; MacGregor, Ormerod, & Chronicle, 2001). Furthermore, research exploring the analogical transfer of information during the solution of insight problems has revealed that transfer in insight is met with the same successes and failures as transfer in other problem-solving situations (e.g., Chen & Daehler, 2000; Ormerod, Chronicle, & MacGregor, 2006). Considering the aforementioned overlap, it makes sense that an affective experience similar to that occurring with an *Aha*, and perhaps other cognitive processes typically associated with insight, could operate when achieving solutions with analysis.

Historically, researchers have identified the distinctive emotional qualities of the insight experience and the mystique or inexplicability of the process as evidence that the cognitive mechanisms involved in achieving insight solutions must be distinct from ordinary noninsight solving. Theorists adhering to this view are in the “special-process” camp, and evidence to support their perspective has steadily accumulated (Anderson et al., 2009; Aziz-Zadeh, Kaplan, & Iacoboni, 2009; Bowden & Jung-Beeman, 2003; Jung-Beeman et al., 2004; Knoblich, Ohlsson, Haider, & Rhenius, 1999; Kounios et al., 2006, 2008; Luo, Niki, & Phillips, 2004a; Mai, Luo, Wu, & Luo, 2004; Schooler & Melcher, 1995; Smith & Kounios, 1996).

Accumulating evidence demonstrating that the right cerebral hemisphere makes a unique contribution to insight not evident in analytic processing has led many researchers to accept that insight is a special process. For example, several studies investigated the time course of hemispheric differences in solution activation for compound remote associate (CRA) problems (Beeman & Bowden, 2000; Bowden & Beeman, 1998). CRAs, adapted from Mednick's (1962) remote associates task, are brief problems in which participants are presented with three words (e.g., CRAB, PINE, SAUCE) and must generate a solution word (e.g., APPLE) that

can be combined with the problem words to yield a compound word or familiar phrase (e.g., CRABAPPLE, PINEAPPLE, and APPLE SAUCE). Like many problems solved with insight, CRA solutions rely on unusual or remote associations, which are likely to capitalize on semantic processing in the right hemisphere. This hypothesized asymmetry stems from the left hemisphere's tendency to mediate fine semantic coding in which a small number of close associates are activated, whereas the right hemisphere mediates coarse semantic coding involving the weak, diffuse, activation of a larger number of distant associates, often required to solve such problems.

In several studies, participants worked on CRAs for various time limits; if they failed to solve a problem before the time limit, they were presented with a target word to name (read aloud). Target words were either solution words or unrelated (solutions to other problems), and they were presented to either the left visual field (right hemisphere) or the right visual field (left hemisphere). Participants named solution words faster than unrelated words, that is, demonstrated solution-related priming for problems they had failed to solve. Participants showed more solution-related priming when naming targets presented to the left visual field (right hemisphere) than targets presented to the right visual field (left hemisphere) for both solved and unsolved problems, and this asymmetry increased with longer solving times (Beeman & Bowden, 2000; Bowden & Beeman, 1998). The right hemisphere advantage in priming was so strong that, when participants were asked to decide whether the words presented were solutions to unsolved problems (yes or no), their responses (button press) were significantly faster for words that were presented to the left visual field (right hemisphere), without sacrificing accuracy for speed. Thus, the typical left hemisphere advantage for responding to words was reversed. Beeman and Bowden's results were in accord with the findings of previous research (Fiore & Schooler, 1998) showing that hints to insight problems are more effective when presented to the left visual field (right hemisphere) than when presented to the right visual field (left hemisphere).

Functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) have also shown lateralized differences between insight and analytic solutions of CRAs. Participants solved problems while brain activity was monitored, and they reported after each solution whether they used

insight or analysis to solve the problem. Participants were told to classify solutions that arrived suddenly, seemed obviously correct, and were achieved without clear intermediate steps as insight solutions, and those that were achieved more methodically through conscious analysis as noninsight solutions. fMRI showed that insight solutions were associated with distinct lateralized activity in the right anterior superior temporal gyrus (STG), activity not evident in noninsight solutions (Jung-Beeman et al., 2004). A follow-up study (Subramaniam, Kounios, Parrish, & Jung-Beeman, 2009) confirmed this right temporal activity, as well as activity in anterior cingulate cortex and parahippocampal areas, which were just below threshold in the original study. In a separate experiment of the original study (Jung-Beeman et al., 2004), high-density EEG revealed a sudden burst of high-frequency gamma-band activity over the right anterior temporal region about 0.3 seconds prior to the button press indicating solution (thus, approximately coinciding with awareness of the solution), which was localized to a region close to that identified in the fMRI experiment. Because the STG is thought to be involved in semantic integration (e.g., St. George, Kutas, Martinez, & Sereno, 1999), and because the right hemisphere has been found to play a significant role in processing distant semantic associations and figurative language (Jung-Beeman, 2005), this activation is thought to be linked to the sudden integration of semantic information resulting in the solution.

In addition, about 1.5 seconds before participants solved problems with insight, a burst of low-alpha EEG activity occurred over the right parietal-occipital cortex (see Fig. 24.1), subsiding just as the right temporal EEG gamma burst began (Jung-Beeman et al., 2004). Low alpha activity over visual cortex is understood to reflect visual sensory gating (cf. Ray & Cole, 1985). The researchers argued that this burst of alpha-band activity signifies a brief deactivation of visual cortex, reflecting a reduction of distracting sensory inputs. Similar posterior alpha activity was measured over this region in an EEG study that explored the restructuring component of insight (Wu, Knoblich, Wei, & Luo, 2009). Wu et al. used a Chinese-character task that required chunk decomposition, a form of restructuring, to break down complex Chinese characters to form new target characters. Their results support the idea that the attenuation of visual inputs facilitates representational change.

Recently, Aziz-Zadeh and colleagues (2009) used fMRI to examine activity associated with the

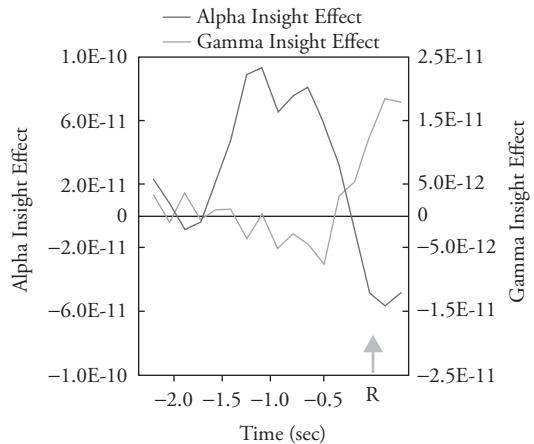


Fig. 24.1 The time course of the insight effect. Electroencephalogram (EEG) alpha power (9.8 Hz at right parietal-occipital electrode) and gamma power (39 Hz at right temporal electrode) for the insight effect (i.e., correct insight solutions minus correct noninsight solutions, in μV^2). The left y-axis shows the magnitude of the alpha insight effect (purple line); the right y-axis applies to the gamma insight effect (green line). The x-axis represents time (in seconds). The yellow arrow and R (at 0.0 s) signify the time of the button-press response indicating that a solution was achieved. Note the transient enhancement of alpha on insight trials (relative to noninsight trials) prior to the gamma burst signifying insight. (Reproduced from open source article by Jung-Beeman et al., 2004.) See color figure.

solution of anagrams by insight, and they found relatively greater right prefrontal cortex (PFC) activity associated with insight versus analytic solutions. Such results build on other recent findings demonstrating right-PFC activity associated with creativity and the production of novel ideas in normal adults (e.g., Howard-Jones, Blakemore, Samuel, Summers, & Claxton, 2005). Moreover, individuals high in schizotypy, who tend to use loose associations processed in the right hemisphere (see Mohr, Graves, Gianotti, Pizzagalli, & Brugger, 2001) solve classic insight problems at a higher rate compared to healthy, nonschizotypal adults (Karimi, Windmann, Güntürkün, & Abraham, 2007), further supporting the critical role of the right hemisphere in insight.

Although there seems to be substantial and increasing evidence that insight processes rely more on the right hemisphere than analytic processes do, such findings should not be interpreted to suggest either that the right hemisphere is exclusively used in insight and not analysis, or that left-hemisphere processing is not needed for insight. Rather, the fMRI and EEG literature indicates that, though the two strategies share many processes, additional activity in right temporal cortex and other areas

(Subramaniam et al., 2009) is present when solvers produce insight solutions. Regardless of the outcome of the debate on the hemispheres in insight, the literature supports the idea that the solution of verbal problems through insight is associated with a unique and sudden integration of weakly and distantly related problem components, and that briefly blocking sensory input facilitates the retrieval into conscious awareness of solution-related ideas initially represented at an unconscious level.

It is possible that other mechanisms, such as the application of unconstrained hypothesis generation, could explain the hemispheric asymmetry associated with insight (e.g., Vartanian & Goel, 2005). In fMRI research with anagrams, Vartanian and Goel observed significant activation in the right ventral lateral prefrontal cortex when participants solved semantically constrained (i.e., anagram letters to be rearranged to generate a word in a specific category) versus unconstrained problems (anagram letters to be rearranged to generate a word with no category specification), which they suggest indicates the role of this region in hypothesis generation in unconstrained situations. Insight problems have been labeled as unconstrained in prior research (e.g., Isen, Daubman, & Nowicki, 1987). Specifically, insight problems often either mislead people trying to solve them, because they imply one solving strategy (or one interpretation or association of problem concepts) or simply lack constraints suggesting the correct approach.

Another distinctive feature of insight is the relative inaccessibility of insight mechanisms to conscious analysis. Research on metacognition and insight has revealed that participants appear to lack conscious awareness of solution-related ideas during intermediate stages in the solving process (Metcalfe & Wiebe, 1987) and are limited in their ability to explain these thoughts aloud (Schooler, Ohlsson, & Brooks, 1993; to be discussed later). The limited accessibility of insight-related processing to conscious awareness may also be linked to the unique contributions of right-hemisphere processing to insight. Research with split-brain patients has demonstrated that conscious experience in these patients with divided hemispheres is more strongly tied to left-hemisphere processing (e.g., Gazzaniga, 1998), whereas right-hemisphere processing tends to influence their behavior without awareness.

Based on the accumulated evidence, especially the brain imaging data, a strong version of the

“business-as-usual” view of insight no longer seems tenable. Insight appears to involve “special” processes.

Insight and Conscious Awareness

When Gestalt psychologists first began to discuss insight about a century ago, they distinguished between insight and analysis based mostly on the different affective experience that accompanies insight but also because insight sometimes comes incidentally when one is not directly focusing on the problem. The Gestalt psychologists’ view was supported not only by shared experience but also by famous anecdotes of scientific insights that led to creative solutions to vexing problems. Some of the most famous of such insights were Archimedes’ eureka moment in the bathtub, Newton’s falling apple, and Poincaré’s bus ride. One thing that these discoveries have in common is that, at least according to the stories, these great thinkers were not consciously considering the problem at the time they experienced insight (see Simonton, Chapter 25). In contrast, since analytic thought is by definition deliberate, solution by incidental analysis is an oxymoron.

Insight problem solving is also different from analytic problem solving in that insight seems to involve processing that renders it inaccessible to metacognition (see McGillivray et al., Chapter 33). Metcalfe and Wiebe (1987) asked participants to periodically judge during the course of problem solving how close they were to achieving a solution. They found that prior to analytic solutions, subjects reported gradually increasing closeness to solution. In contrast, prior to insight solutions, subjects reported little or no progress toward solution right up until the point in time at which solution was imminent (cf., Smith & Kounios, 1996). Metcalfe and Wiebe, therefore, concluded that insight was fundamentally a different process, one that involves critical mechanisms that go on outside awareness.

Using a somewhat different paradigm in which participants were interrupted during solution and asked to give a verbal description of their solution efforts, Schooler and Melcher (1995) found a difference in conscious access to solution information between insight and analytic problem solving. Schooler and colleagues (1993) found that verbalization during solving interfered with the solution of insight problems but not analytic problems. Based on these findings, Schooler proposed a verbal overshadowing effect on insight due to the

inaccessibility of critical insight processes to verbal description. A recent investigation by Gilhooly, Fioratou, and Henretty (2010) suggests that verbal overshadowing might be specific to visual problems, regardless of insight or analytic solution strategy. Nevertheless, much of the research on metacognition and verbal overshadowing of insight suggests that steps in the insight process are beyond conscious analysis. Furthermore, it is possible that the inaccessibility of insight processes to linguistic analysis may be due to the involvement of the less-verbal right hemisphere.

The unconscious processing of insight antecedents may be a critical factor in the eventual *Aha!* reaction that accompanies insight solutions. The sudden conscious awareness of the solution can be particularly surprising given a comparison to analytic processes that may even be “reasoned out” either vocally or subvocally as the solver advances toward completion (Newell & Simon, 1972; Smith & Kounios, 1996; Thorndike, 1898).

Knowledge Selection and Fixation

In the first half of the 20th century, the Gestalt psychologists proposed that insight involves the application of a special type of knowledge that is different from that used in trial-and-error problem-solving strategies (e.g., Duncker, 1945; Koffka, 1935; Köhler, 1925; Wertheimer, 1945/1959). They argued that it was the incorrect application of prior knowledge that prevented the achievement of insight and that insight is facilitated only when problem solvers go beyond trial-and-error processing to acquire extraordinary knowledge structures (e.g., Duncker, Köhler). Ordinary thought is *reproductive*, involving the reuse or adaptation of older ideas or approaches, whereas insight requires *productive thought* (Wertheimer, 1945/1959; see also Mandler & Mandler, 1964) in which a deeper conceptual understanding allows problem solvers to select relevant knowledge components and combine them in novel ways, or to guide problem solvers to attend to the environmental stimuli that are most relevant. Wertheimer suggested that insight was not the result of the problem solver blindly recombining problem elements in search of a solution; rather, he argued that it requires that the problem solver gain the necessary structure on which to build a solution. Thus, a lack of success in problem solving could stem either from the retrieval of irrelevant information from long-term memory or the retrieval of relevant components that are applied or linked within an inappropriate structure.

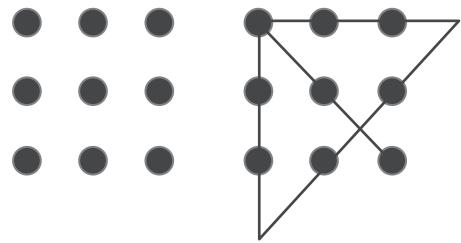


Fig. 24.2 The nine-dot problem. (*Left*) Subjects' task is to draw four straight lines that go through all nine dots without backtracking or lifting the pencil from the paper. (*Right*) One solution to the problem.

Problem solvers often do poorly in solving classic insight problems, such as the nine-dot problem (Maier, 1930; see Fig. 24.2). Researchers have suggested that participants fail to solve this problem (solution rates are consistently less than 10% without hints) because either they do not possess solution-relevant knowledge about extending lines “outside the box,” or because they lack the visuospatial capacities to mentally construct the solution configuration (e.g., Chronicle, Ormerod, & MacGregor, 2001; Kershaw & Ohlsson, 2004). However, when experimenters provide participants with training problems and strategies and give hints about the solution (e.g., Lung & Dominowski, 1985; Weisberg & Alba, 1981) or provide perceptual hints about the overall solution configuration (Chronicle et al. 2001; MacGregor et al., 2001), problem solvers achieve only modest gains in solution rates. Kershaw and Ohlsson had to provide extensive knowledge about pivoting solution lines on non-dot points and to train participants to think outside the constraints of the problem presentation space in order to substantially facilitate solution rates.

The limited transfer of problem-related information during the solution of the nine-dot problem mirrors results observed in prior research on analogical transfer using Duncker's (1945) radiation problem (see Holyoak, 2005, for a review). In the radiation problem, another classic insight problem, participants are asked to generate a solution that will enable an inoperable stomach tumor to be eradicated by allowing enough radiation to reach the tumor without destroying the surrounding healthy tissue. The problem can be solved by employing the convergence principle: Multiple rays of weaker intensity can be applied from various positions converging with sufficient strength at the tumor. In preliminary research on transfer, participants were

provided with one or more source problems that also relied on the convergence principle for solution (Gick & Holyoak, 1980, 1983). When a relevant source problem was provided prior to the radiation problem, few participants transferred the convergence principle solution unless they were explicitly told that the source information might be helpful in generating the solution. The application of the solution principle was more likely if multiple source problems with convergence solutions were presented for comparison instead of one (Gick & Holyoak, 1983; Kurtz & Loewenstein, 2007), when the source problem appeared similar in character and theme (e.g., a patient with a brain tumor) to the target problem than if the source problem was from an unrelated domain (Keane, 1987), or when abstract statements explicating the solution principle were provided (Chen & Daehler, 2000). These findings suggest that gaining problem-relevant knowledge may do little to facilitate solution success in insight unless that knowledge results in the acquisition of a deep conceptual understanding of the problem's components.

The Gestalt psychologists felt that repeated application of incorrectly selected knowledge from long-term memory (i.e., fixation) could prevent the deep conceptual understanding necessary to achieve insight (Duncker, 1945; Luchins, 1942; Maier, 1930; Scheerer, 1963). Duncker conducted much of the pioneering work on fixation, most notably with the candle problem. In the candle problem, researchers present participants with a candle, a book of matches, and a box of tacks, and ask them to devise a way to attach the candle to a door so it can burn properly. The insightful solution to this problem is to empty the tack box and use it as a ledge to support the candle. However, when problem solvers first approach the problem and see the box filled with tacks, they become functionally fixated on the standard function of the box as a container, which limits their ability to consider the tack box as having other functions that may potentially facilitate the solution to the problem. Perseveration on an object's dominant function may be so persistent that the solver reaches an impasse and becomes increasingly fixated on a solution idea, repeatedly attempting the same approach (Smith, 1995).

Modern theorists have suggested that when a problem solver's initial problem space contains irrelevant or incorrectly constructed prior knowledge, the problem solver will reach an impasse that prevents further constructive work on the problem

(Knoblich et al., 1999; Ohlsson, 1992). Knoblich and colleagues proposed that during impasse, a problem solver's prior knowledge leads him or her to incorrectly apply constraints to the problem-solving situation that limit the possibilities for solution. In their research on matchstick arithmetic problems (see Fig. 24.3), participants' prior knowledge of algebra (which is relevant in most equation solving but is irrelevant in matchstick arithmetic) blocked their consideration of relevant ideas during problem solving. For example, participants may have assumed that what happens on one side of the equation must happen on the other side, or that the operands could be subdivided to correct the equation, whereas the operators could not. Continued progress was only possible when these mental constraints were relaxed and additional ideas were considered for solution.

In imaging research exploring comprehension in insight, Lang and colleagues (2006) recorded event-related brain potentials (ERPs) in participants who gained an insightful understanding of the underlying structure in the number reduction task (NRT) and compared them to ERPs of those who did not achieve such an understanding. In the NRT, participants apply rules to sequentially presented numbers to produce a new string of numbers, the last of which is most important. For all trials, the pattern of cues was *ABCCB*, such that the last answer is the same as the second answer. Participants who came to explicitly understand that they could skip to the end of the sequence based on this pattern were viewed as achieving insight. Lang and colleagues reported that those who eventually achieved insight had greater slow-positive-waveform (SPW) amplitude across the length of each trial over parietal electrode sites and a relatively enhanced P3a

$$\text{Problem: IV} = \text{III} + \text{III}$$

$$\text{Solution: VI} = \text{III} + \text{III}$$

$$\text{Problem: III} = \text{III} + \text{III}$$

$$\text{Solution: III} = \text{III} = \text{III}$$

Fig. 24.3 Sample matchstick arithmetic problems (after Knoblich et al., 1999, p. 1536). The goal is to move a single stick in such a way that the initial false statement is transformed into a true arithmetic statement.

component over frontal-central sites, both observed from the outset of the task, compared to those who failed to achieve insight. Researchers had previously tied such effects to increased working-memory involvement (Vos et al., 2001) and the perceived novelty or distinctiveness of the stimuli (P3a; Gaeta et al., 2003). Considered together, these results suggest that thought processes applied in earlier trials determined whether participants would absorb the correct knowledge to achieve a deep conceptual understanding of the task.

Restructuring

Insightful problem solving seems to depend to a significant extent on the problem solver making one or a combination of three basic mistakes: the solver misrepresents the problem elements in such a way as to preclude solution; the solver focuses on information retrieved from memory that is not pertinent to obtaining the solution and may in fact lead him or her away from the solution (Knoblich et al., 1999; Ohlsson, 1992; Seifert, Meyer, Davidson, Patalano, & Yaniv, 1995); or the solver works with insufficient information to achieve success (Kaplan & Simon, 1990). Because they have made one of these critical errors preventing them from reaching solution, the solvers' focus on the unproductive line of reasoning must be broken via interruption or restructuring of thought allowing a shift in solution strategy (Ohlsson, 1992; Weisberg, 1995) and new paths to solution.

There are several theories as to how we change representations or restructure a problem. Gestalt psychologists stressed the importance of restructuring in insight processing and described it as an automatic process that occurs as you attempt (and fail) to solve the problem—you simply “see it in a new light” (Duncker, 1945; Koffka, 1935; Köhler, 1925). According to the cognitive view (Ohlsson, 1992), problem solvers develop a representation of the problem and apply heuristics to transform the problem space so that it looks like the solution space. Eventually, when progress stops, they apply the “restructure when stuck” heuristic (Kaplan & Simon, 1990).

As described previously, people experiencing insight are usually unable to report the processes by which they are able to restructure a problem or even that they did restructure a problem. Therefore, to explain the insight experience, restructuring theories must make room for unconscious reorganization of unsuccessful knowledge structures in favor of continued progress, though not necessarily in

a series of increments, as is common in analysis. Knoblich and colleagues (1999; see also Ohlsson, 1992) have suggested just such a process in their *representational change theory*. First, the problem solver pursues and rejects nearly all the possibilities within the presently available problem space. Once at impasse, the problem solver can reject assumptions about the problem that were made based on incorrectly selected prior knowledge, thereby clearing the path for exploring a problem space based on different assumptions. Finally, reorganization and restructuring are spontaneous and out of conscious control. For example, after reaching impasse when attempting to solve a matchstick arithmetic problem (Fig. 24.3), the participant may refocus on a non-numeric aspect of the problem that is also depicted by the matchsticks, such as the addition sign, as an element of the problem space that may be altered.

When Chronicle, Ormerod, and MacGregor proposed their *progress monitoring theory* (2001; MacGregor et al., 2001; Ormerod, MacGregor, & Chronicle, 2002), they suggested that fixation and impasse may occur prior to insight but that they are not necessary to the process. According to their theory, the problem solver actively monitors progress toward solution and is continually applying problem-solving heuristics that enable him or her to change strategy based on a lack of progress toward solution. Fleck and Weisberg (2004) generated support for the progress-monitoring theory in research using the verbal protocols of solvers attempting Duncker's candle problem. They found that participants often restructured their assumptions about the problem after they rejected solution ideas as implausible without actually having to implement them.

Other researchers have been skeptical that restructuring has to be sudden or that it even has to occur outside of awareness. In an early study on failure and insight, Weisberg and Suls (1973) proposed that failed solutions during the solving process allow participants to acquire additional knowledge regarding problem components that alters future solution attempts. For example, in the candle problem, participants may have discovered that the tacks were too short to go directly through the candle. They may then have wondered what the tacks could penetrate and subsequently made the leap to tacking through the relatively thin box. A solution that could seem like a release from functional fixedness with regard to the nature of the box may have actually been a logical extension of knowledge gained based on previous failures.

Durso and colleagues (1994) reported that participants gradually accumulated solution-related knowledge when they solved anagrams, rather than experiencing sudden, all-or-none solutions. Durso and colleagues asked participants to rate the similarity of word pairs to infer the location of words within the semantic network and to demonstrate the change in network structure that accompanies restructuring. Beginning before problem presentation, continuing during the solving process, and even after solution, participants gradually rated words that were critically involved in the restructuring process as being more similar, showing that, at some level, they may have been steadily accumulating the necessary solution-related information needed to take the critical step of restructuring.

In a series of experiments using fMRI, Luo et al. (2004a) explored the neural correlates of restructuring. It is important to note that in their experiments participants did not solve problems, but rather recognized solutions (or cues), so the cognitive and neural processes necessarily would be expected to differ from those involved in pure solving. Indeed, the patterns of brain activity observed (Luo et al. 2004a, b; Mai et al. 2004) overlap with, but differ from, those observed when people solve problems (Jung-Beeman et al., 2004; Kounios et al., 2006; Subramaniam et al., 2009). Luo et al (2004a) had 13 subjects read incomprehensible sentences followed by solution cues that would eventually trigger an alternative interpretation of a concept that was critical to understanding the sentence, e.g., "You could not tell who it was, because a professional took the photo of that old man (x-rays)." Or "His position went up because his partner's position went down (see-saw)." Participants were presented with a sentence for 7 seconds and asked whether they understood it. Then they were shown the response cue and asked if they understood it in the new context. Participants were assumed to have achieved insight if they initially failed to understand the sentence but understood it after the cue. They found a correlation of insight with activity in anterior cingulate cortex, an area known to monitor cognitive conflict (Carter et al., 2000), and left lateral prefrontal cortex, an area thought to select semantic representations from among competing alternatives (Thompson-Schill, D'Esposito, Aguirre, & Farah, 1997). In addition, the level of ACC activation present in insight trials decreased across blocks, suggesting that the ACC is more involved when the task is novel, before participants develop strategies that facilitate comprehension of sentences of this type.

In similar research, Mai, Luo, and colleagues (2004) presented participants with answers to riddles that they could not solve and examined their neural activity with fMRI. They proposed that when participants switched from their initial, prepotent response to a representation that coincides with the riddle's solution (a process possibly similar to restructuring), this should initiate cognitive conflict reflected by increased activation in the anterior cingulate cortex (ACC). Using ERPs time-locked to the solution word's presentation, Mai and colleagues reported greater negativity for insight trials than noninsight trials 250–500 ms post solution, localized via dipole modeling to the ACC.

In addition, Luo and colleagues (Luo, Niki, & Phillips, 2004b, as cited in Luo & Knoblich, 2007) observed greater ACC activation when participants attempted to solve riddles with heterogeneous solution types than when participants were presented with riddles of a single solution type. Therefore, the role of the ACC in this paradigm may be to facilitate abandonment of incorrect problem representations. However, as the novelty of the problem-solving environment decreases, there may be less experience of cognitive conflict, so the ACC may not need to mediate conflict as actively. It should be emphasized again, though, that these researchers were likely studying a representational change phenomenon that differs from insight as classically conceived—more of a recognition or understanding moment than a generative insight—although the two processes may involve some similar mechanisms.

To what degree is restructuring determined by more fundamental cognitive abilities? While working memory, vocabulary size, and visuospatial ability may help some solvers with aspects of more complex problems, there is little evidence at the moment that they correlate with the ability to restructure a problem, which is the basis of insight. Ash and Wiley (2006) examined the relationship between working-memory capacity and the size of the initial search space of a problem. Working memory is essential for maintaining the elements of a problem during solving, especially for problems with large search spaces (i.e., problems that require the exhaustion of several approaches prior to restructuring). The researchers compared problems with search spaces of different sizes so they could isolate the restructuring component of insight from other processing that may be necessary as the initial problem space becomes more complex. Ash and Wiley found a relationship between working-memory capacity and

the tendency to restructure, but only for problems with large search spaces. For problems with small search spaces, which do not involve much systematic search, working-memory capacity did not predict restructuring (see also Fleck, 2008; Gilhooly & Murphy, 2005). Thus, when restructuring is isolated from other components of problem solving, individual differences in controlled search and the conscious application of strategies are apparently unrelated to success at solving.

Impasse and Incubation

There is some controversy (Dominowski & Dallob, 1995; Smith, 1995; Weisberg, 2006) about what drives restructuring of the problem space and to what degree impasse is necessary for restructuring. Taking a break from the problem after impasse, a period known as incubation, can promote the solution of insight problems (Christensen & Schunn, 2005; Segal, 2004). Because restructuring is so critical if solvers are to overcome an initial misleading solution idea in favor of novel solution ideas (e.g., Martindale, 1995; Metcalfe & Wiebe, 1987; Ohlsson, 1992), it is vital to understand what scenarios promote restructuring. There is a significant literature on the importance of incubation in this regard.

Various theories have been proposed regarding the nature of incubation and the mechanisms of its effect (see Sio & Ormerod, 2008, for a review). First, researchers have posited that incubation may lead to insight because time away from the problem results in the deactivation of incorrect knowledge representations in the brain (i.e., selective forgetting; Smith & Blankenship, 1991). It has also been proposed that incubation may result in success because the unconscious remains at work (perhaps as spreading-of-activation) while the conscious mind is engaged on another task (e.g., Wallas, 1926; Yaniv & Meyer, 1987). In research testing these theories, Segal (2004) reported that solution rates for insight problems improved when participants took a break after impasse, regardless of the duration. Moreover, this break was more successful when the break interval was filled with a cognitively demanding task. In incubation research conducted by Kohn and Smith (2009), an incubation effect was observed when the to-be-solved problems were completely removed from sight during the incubation period, and not when they remained partially in view, implicating the value of distraction/attention switches in incubation, rather than merely selective forgetting. However, Kohn and Smith (see also Vul & Pashler,

2007) only revealed an incubation effect when participants were exposed to misleading information at a problem's outset, lending support to the selective forgetting hypothesis.

Some researchers (Fleck & Weisberg, 2004; MacGregor et al., 2001) have speculated that restructuring may occur as the result of an internal or external search for new information following impasse or failure. Spontaneous restructuring stemming from exposure to relevant external hints has also received empirical support from research demonstrating that exposure to problem-relevant information during incubation periods can facilitate the insight experience (Maier, 1930; Seifert et al., 1995).

Despite the prominent place of incubation as a means of overcoming fixation in many insight theories, research on fixation and impasse has been sparse in part because it is difficult to operationally define fixation: How long must it last? What is considered "limited" progress? Must there be repeated failure with the same incorrect solution? and so on. Although problems that typically produce insight solutions are often created by inserting misleading components designed to generate fixation on irrelevant aspects (Weisberg, 1995), it can be difficult to tell if solvers have actually been misled. Researchers have had some success verifying and evaluating fixation and impasse by collecting verbal protocols (Fleck & Weisberg, 2004; Kaplan & Simon, 1990) and using eye tracking (Grant & Spivey, 2003; Jones, 2003; Knoblich, Ohlsson, & Raney, 2001; Thomas & Lleras, 2009a). Fleck and Weisberg (2004) examined verbal protocols for the presence of impasse characteristics and were able to demonstrate that impasse-like characteristics were present in the thought processes of approximately 45% of participants attempting to solve Duncker's (1945) candle problem. Statements that led Fleck and Weisberg to classify a participant as being at impasse mostly reflected confusion about specific problem components or an inability to generate additional ideas. However, verbal-protocol analysis is not beyond criticism, because it is based on the assumption that what participants actually verbalize accurately reflects their cognitive processes and that they choose to verbalize all the mechanisms to which they can consciously attend.

The use of eye-tracking technology to measure gaze fixation has shown promise as a method for measuring what may be fixation during problem solving (e.g., Grant & Spivey, 2003; Jones, 2003).

Knoblich and colleagues (2001) operationally defined increased gaze-fixation duration to be indicative of impasse during problem solving, and they speculated that increased gaze fixation on specific problem elements could be reflective of fixation on incorrect problem components. Mean gaze-fixation duration increased across the solving period as participants attempted to solve matchstick arithmetic problems (see Fig. 24.3), indicating that solvers began to reach impasse as time spent working on the problem progressed. Furthermore, solvers fixed their gaze on irrelevant elements of the problem during the initial solving period and fixation on relevant items increased with time. Using a somewhat different approach, Jones (2003) operationally defined the point that impasse occurred in the solving process as the instant gaze fixation was at least two standard deviations above mean gaze fixation for the participant. Jones found that all solvers experienced impasse before they reached solution and speculated that impasse preceded a representational change that led to solution. Thus, eye-movement and verbal-protocol research suggests that fixation and impasse play important roles in the insight process and may be fundamental in establishing the necessary environment for representational change.

Insight: Finding the Proper State of Mind

In addition to incubation, there are other processes that facilitate insight. One of these is positive mood. Researchers have often noted a link between positive affect and creativity (e.g., Amabile, Barsade, Mueller, & Staw, 2005), and several studies have directly examined the role of affect in insight (e.g., Isen et al., 1987; Subramaniam et al., 2009). Isen and colleagues were among the first researchers to demonstrate the relationship between insight and positive affect when they observed that participants' solution rates for Duncker's (1945) candle problem increased when brief comedic films were used to induce positive affect. More recently, Subramaniam and colleagues asked participants to indicate their solution strategy (either insight or analytic) on a trial-by-trial basis for solved CRAs during fMRI scanning and to complete self-report measures of affect and anxiety. When participants were high in positive affect and low in anxiety, they solved more problems overall, especially more problems by insight.

Positive affect may promote insight and creativity because it allows people to broaden attention, both perceptually and conceptually, and consider ideas for solution that would typically fall outside the scope

of their awareness (Rowe, Hirsch, & Anderson, 2007). In research exploring the link between cognitive processes and mood, Rowe et al. observed changes in selective attention and semantic access with positive mood. A positive mood state was associated with significantly weaker selective attention (i.e., a broader scope of attention or a leakier filter) on a flanker task, as well as increased semantic access to remote associations in Mednick's (1962) RAT, than were negative or neutral moods. Mood-based performance on tasks of attention and semantic access was highly correlated ($r = .49$), a result not observed for task performance during neutral or negative moods. Rowe et al. interpreted these findings as further support for a common origin (i.e., an affective influence) for the two processes.

In similar research, researchers who compared attentional resources in creative versus less creative individuals found support for a relationship between attention and creativity. As measured by scores on Mednick's (1962) RAT, those high in creativity tended to use hints they had been instructed to disregard (Ansburg & Hill, 2003). A functional-neuroanatomical explanation for such a relationship may be seen in neuroimaging results reported by Subramaniam and colleagues (2009) in which greater positive mood assessed at the start of the experiment session was associated with an increase in insight solutions, as well as brain-related activation in the ACC, both before and during problem solving. Specifically, level of positive affect modulated dorsal ACC activation (the more positive participants were, the higher ACC activation was) during the preparatory interval prior to all problems that were eventually solved. Across all participants, this preparatory activation was stronger for problems that were subsequently solved with insight than for problems subsequently solved by analysis (see also Kounios et al., 2006). These findings led the researchers to suggest that a positive mood may create a brain state helpful for achieving insight, perhaps by modulating the cognitive control system to better detect (or switch to) weak brain activity associated with more distant associations.

The efficacy of brief training intervals in enhancing insight problem solving (e.g., Ansburg & Dominowski, 2000; Cunningham & MacGregor, 2008; Dow & Mayer, 2004; Lung & Dominowski, 1985; Maier, 1933; Schwert, 2007) has also been tested. Training typically involved advice on avoiding common obstacles to achieving insight: initial ideas during problem solving are often misleading;

problem solving may be difficult because you are applying unnecessary constraints to the problem (see Ansburg & Dominowski, 2000; Cunningham & MacGregor, 2008). After learning the heuristics, participants usually worked on example problems that demonstrated the solution and its associated logic. Results of such training regimes have so far been generally limited to domain-specific or even problem-specific training effects, similar to those present with training to solve other types of problems (Ansburg & Dominowski, 2000; Dow & Mayer, 2004; Gick & Holyoak, 1980). Furthermore, training was most beneficial in enhancing solving ability for traditional insight problems, such as riddles and puzzles, and less influential when the insight problems involved real-life contextual information (Cunningham & MacGregor, 2008).

Investigators have also enhanced insight rates by altering the solving process through provision of implicit hints (e.g., Grant & Spivey, 2003; Sio & Ormerod, 2009; Thomas & Lleras, 2007). Furthermore, Christensen and Schunn (2005) observed an increase in insightful problem solving when external hints related to the solutions were incidentally provided during incubation periods that occurred at regular intervals throughout the problem-solving task.

Hints diverting attention from one problem component to another can facilitate the occurrence of insight (e.g., Grant & Spivey, 2003; Kaplan & Simon, 1990; Thomas & Lleras, 2009a). Grant and Spivey were able to enhance solution rates for Duncker's (1945) radiation problem, described earlier, by presenting the problem with the outside surface of the body flashing, drawing attention to a component that could trigger insight. Thomas and Lleras (2007) also directed participants' attention to the problem's solution by having participants track a series of letters and digits that appeared on the computer screen in a sequence such that the eye-movement pattern mirrored the layout of the problem's solution (i.e., across the surface of the skin and moving in toward the tumor). Similar effects occurred when participants directed their attention to letter and number locations resembling the solution pattern without physically moving their eyes to track the stimuli (Thomas & Lleras, 2009a). Furthermore, having participants engage in physical movements that coincided with the solution to Maier's (1931) two-string problem (i.e., swinging their arms back and forth in a pendulum-like motion) facilitated problem solving (Thomas & Lleras, 2009b). This

research collectively supports the benefit of shifting attention from misleading to relevant problem components in facilitating insight.

In addition to facilitatory effects from external hints, the likelihood of experiencing insight has been associated with specific preparatory brain states. For example, Kounios et al. (2006) discovered that neural activity immediately prior to the presentation of a problem predicts whether solutions will arise with insight or analytically. They asked participants to solve a series of CRA problems using the insight judgment procedure. Of interest was the time window preceding the presentation of a CRA problem. EEG and fMRI measures of neural activity during this preparatory time window predicted whether the subsequently displayed problem would be solved with insight or analytically. The analysis of low-alpha EEG activity (8–10 Hz) revealed greater preparatory activity for trials subsequently solved with insight measured over midfrontal, bilateral temporal, and bilateral somatosensory regions. In contrast, the interval preceding trials solved analytically was associated with greater activity over posterior brain regions. A parallel experiment with fMRI generally replicated the results of the EEG experiment and further clarified the neural components underlying the strategy differences, identifying increased signal strength for insight trials in the ACC, posterior cingulate cortex, and bilateral middle and superior temporal gyri. These results suggest that a form of preparation before problem presentation helps determine whether a subject will tackle the subsequent problem with an insight or analytic strategy. Preparation for insight apparently involves inward focus of attention and priming of brain regions supporting semantic processing. In contrast, preparation for analysis involves outward focus of attention on the screen on which the expected problem will appear.

A more recent study by the same researchers revealed that even resting-state brain activity, that is, brain activity when an individual is not engaged in task-directed cognition, predicts that individual's likelihood of later solving problems insightfully or analytically. Kounios and colleagues (2008) grouped participants into high-insight and high-analytic groups based on the proportions of anagrams they solved with insight. High-insight subjects showed greater right-hemisphere EEG activity and more diffuse activity of visual cortex suggestive of broader attention in accord with previous research highlighting an association among insight, right-hemisphere activity, and broad attention (Fiore & Schooler,

1998; Jung-Beeman, 2005; Martindale, 1995). These results demonstrate that the tendency to solve problems with one or the other strategy is a function of states and processes that begin well before the presentation of a problem. Indeed, these states may be relatively stable and constitute a dispositional cognitive style. Such studies of insight-related preparatory and resting-state activity imply that getting the problem solver into the appropriate brain state may be one of the most effective means to enhance insight.

Conclusions: Current Perspectives in Insight

In the current chapter we reviewed the themes and advances in insight research over the last century. Regarding the debate between “business-as-usual” and “special-process” views of insight, research revealing right-hemisphere contributions to insight not present in analysis, as well as the contribution of unconscious processing in insight, has strengthened the perspective that insight is fundamentally different from analysis. Insight researchers have expanded our understanding of the components of insight to include the occurrence of impasse/fixation and the eventual restructuring of thought. Much of the research surrounding these components has focused on mechanisms that facilitate restructuring, and therefore, insight, including the benefits of incubation, external hints, and preparatory mind states. Next we review three of the most significant developments in insight research since the resurgence of research in this area.

Three recent developments in insight research have substantially changed how we study insight and what we know about its cognitive, neural, and affective underpinnings. First, methodological developments in the construction of problem stimuli have circumvented criticisms of the traditional approach of studying insight based on classic insight problems. These methodological developments have laid the groundwork for the second development, namely, the application of neuroimaging techniques to study aspects of insight that could not easily be studied using traditional methods (see Morrison & Knowlton, Chapter 6). Third, there is a new emphasis on examining factors that enhance insight. These three points are briefly discussed next.

Although much was learned in the decades spent studying the solution of more complex “classic” insight problems such as the candle, nine-dot, and two-string problems, more recent innovations

in stimulus construction and methodology have begun to transform the study of insight. The traditional approach of comparing performance on such insight problems with performance on analytic problems has two basic flaws. First, while *processing* can be insightful or analytic, *problems* are neither. Generally, classic insight problems are often solved with a flash of insight. But there is nothing about such problems that *requires* that they be solved with insight. Undoubtedly, subjects sometimes solve them analytically. So the assumption that so-called insight problems are always solved insightfully is problematic. Second, the analytic and classic insight problems used in past studies are complex. Evidence that different types of processes are used to solve them is ambiguous because these two classes of problems differ from each other in many ways—working memory load, modality specificity, and so on—and not just in terms of the insight/analytic distinction.

One approach to overcoming this problem is the development of larger sets of smaller, relatively homogeneous, problems amenable to standardization and norming. An important example of this approach is the development and application of CRA problems (described earlier). These problems can be solved within a few seconds, allowing researchers to acquire more data per subject within a session, which is a prerequisite for neuroimaging studies (see Jung-Beeman et al., 2004, for more on the advantages of such stimuli). Though it is possible that aspects of the insight process are lost when simpler stimuli are utilized, distinctions between the insight processes involved in the solution of classic insight problems and such modern stimuli, should they exist, will be revealed as research in the field progresses.

A related development, based on the notion that a defining feature of insight is the sudden awareness of the solution, is the systematic use of subjects’ judgments about their own solving strategies. As described earlier, the studies of Beeman, Bowden, Kounios, and others require subjects to report, for each solution, whether it was the product of insight or analysis, a distinction with which virtually all subjects indicate that they are familiar. In this way, insightful and analytic processing can be directly compared while controlling for ancillary differences between the stimuli that result in these two types of processes by using many examples of a single type of problem (e.g., CRAs and anagrams).

These methodological developments have fueled an ongoing series of neuroimaging studies of insight

focusing on aspects of problem solving that are, at best, difficult to study using traditional behavioral techniques. As noted earlier, researchers using fMRI and EEG have corroborated the unique role of right-hemisphere processing in insight (Aziz-Zadeh et al., 2009; Jung-Beeman et al., 2004), observed initially in behavioral research. Furthermore, this research has clarified the role of sensory gating and semantic integration in the solution of verbal problems with insight (see Jung-Beeman et al., 2004). Researchers have also linked activity in the ACC, a region associated with conflict monitoring, to the restructuring component of insight (Luo et al., 2004a; Mai et al., 2004); this activation may signify processing involved in the abandonment of incorrect problem representations in favor of continued progress. Finally, neuroimaging has contributed to our understanding of the resting (Kounios et al., 2008) and preparatory (Kounios et al., 2006) brain states associated with the subsequent occurrence of insight.

Because insight can be instrumental to real-world innovation, researchers have also been expanding their efforts to isolate components of insight and factors that influence or facilitate these components. Studies have shown that some components of insight processing rely on the same core abilities as analytic processing, such as working memory, fluid intelligence, general problem-solving ability, and vocabulary (Ash & Wiley, 2006; Davidson, 1995; Fleck, 2008; Gilhooly & Murphy, 2005; Schooler & Melcher, 1995; Sternberg & Davidson, 1982), so methods that improve more fundamental problem-solving abilities (see Koedinger & Roll, Chapter 40) would naturally be expected to improve insight as well. Researchers have also begun to explore correlates of insight that could be exploited in the future as mechanisms for enhancing insight. For example, one correlate that investigators have identified is divergent thinking (Ansburg, 2000; Davidson, 1995; DeYoung, Flanders, & Peterson, 2008). Divergent thinking is the ability to rapidly generate multiple solutions for a single problem, such as listing as many uses as possible for a brick. Divergent thinking has been explored extensively in creativity research (Guilford, 1950) and efforts have been made to enhance it (Clapham, 2001; Runco & Okuda, 1991).

The aforementioned developments in problem stimuli and neuroimaging methods, as well as additional knowledge regarding insight components and factors relevant in facilitating insight, have set the stage for further advances in insight research. We conclude with a discussion of some of the questions

and concerns to be considered by researchers in this advancing field.

Future Directions

The evolution of theories and methodologies in the study of insight has fashioned an ideal climate to enable researchers to continue the exploration of fruitful veins of research, as well as consider topics that have received little attention to date. Neuroimaging offers continued promise in isolating and identifying components of insight, such as restructuring and impasse. We believe that employing a broad range of advances in neuroscientific techniques and theories constitutes an important approach to further elucidating mechanisms of insight and adjudicating among contrasting theories. However, to reach their full potential, neuroscientific approaches must be fully integrated, both in terms of methodology and theory, with behavioral techniques and cognitive theory. For example, metacognitive research supports the notion of sudden conscious awareness of an insight near the point of solution (Metcalfe & Wiebe, 1987; Smith & Kounios, 1996), whereas other behavioral research indicates the accumulation of partial information prior to solution (Durso et al., 1994). If both veins of research are valid and the possibly gradual reorganization of thought occurs outside of awareness and enters consciousness in a sudden leap, then we should be able to determine the point at which enough solution-relevant information has been acquired or activated on the unconscious level to enter into conscious awareness. This should be possible by integrating existing behavioral methods with neuroimaging techniques, thus permitting researchers to correlate the conscious experiences of the individual with associated neural activity.

Though neuroimaging studies have begun to isolate insight-related brain regions thought to reflect relevant information-processing components, equally important is the identification of component sequencing and mutual influence exemplified, for example, by how aspects of impasse may affect restructuring (Fleck & Weisberg, 2004). If insight is best conceptualized in terms of constraint satisfaction implemented by parallel processing, then the use of neuroimaging techniques should be expanded beyond identifying critical brain regions to explore how these regions work together to yield insights (e.g., Payne & Kounios, 2009).

While we learn more about insight itself, we must continue to explore how insight is related to other

forms of creativity and innovation (see Smith & Ward, Chapter 23). Although hypothesized connections among insight and other forms of creativity are helping to stimulate interest in insight, research examining the hypothesized overlap between insight and other types of creative cognition is lacking. Such research will contribute toward delineating the field of creativity research, which suffers from vague and outmoded definitions.

Clarifying the relationships among insight and other forms of creativity is particularly important for the development of techniques to enhance insight (e.g., Cunningham & MacGregor, 2008). Research in the field of creativity enhancement has achieved recent successes, which may very well be generalizable to insight. For example, Markman, Lindberg, Kray, and Galinsky (2007) have evaluated the use of *counterfactual mindsets* as a means of enhancing either creative or analytic thought. In their research, priming people with additive counterfactuals (achieved through statements modifying reality by adding elements to a situation) induced a mindset that enhanced creativity, whereas priming with subtractive counterfactuals (achieved through statements modifying reality by removing elements from a situation) enhanced analytic thought. In a similar manner, Friedman and Förster (2005) have successfully enhanced either creative or analytic problem-solving success by inducing approach or avoidance motivational states. If the relationship between insight and other forms of creativity can be clarified, then these and other creativity-enhancement factors and techniques can be examined for their potential efficacy in facilitating insight.

Interest in facilitating insight, combined with recent advances in delineating the cognitive components and functional neuroanatomy of insight, has raised the possibility of using brain-stimulation techniques such as transcranial direct current stimulation or transcranial alternating current stimulation to enhance insight during problem solving (van Steenburgh, 2011). This approach will undoubtedly contribute to our understanding of the neural basis of insight, though whether brain stimulation ever becomes a practical technique for enhancing insight is currently unknown.

Recent advances in insight research have fostered additional questions for future research to answer. For example, to what degree are the existing research findings generalizable? By defining insight in terms of the subjective sudden awareness of a solution or interpretation, are we lumping together phenomena

that should really be considered separate? Is recognition that a presented solution is correct ("Uh-duh;" Luo et al., 2004a; Mai et al., 2004; Qiu et al., 2008) the same as generating an insight solution oneself ("Aha;" Jung-Beeman et al., 2004; Knoblich et al., 1999)? To what degree can we make conclusions about perceptual insight based on findings using primarily verbal stimuli (Gilhooly et al., 2010)? When an individual solves a CRA problem, an anagram, or, for that matter, a classic insight problem, is this similar to what happens when a person makes a scientific breakthrough (see Dunbar & Klahr, Chapter 35) or comes to a new realization about a real-world situation? With new methodological tools and theoretical perspectives, researchers must take advantage of the current momentum in insight research to broaden their efforts in ways that address such concerns and bridge the gap between the laboratory and life.

References

- Amabile, T. M., Barsade, S. G., Mueller, J. S., & Staw, B. M. (2005). Affect and creativity at work. *Administrative Science Quarterly*, 50, 367–403.
- Anderson, J. R., Anderson, J. F., Ferris, J. L., Fincham, J. M., & Jung, K-J. (2009). Lateral inferior prefrontal cortex and anterior cingulate cortex are engaged at different stages in the solution of insight problems. *Proceedings of the National Academy of Science USA*, 106, 10799–10804.
- Ansburg, P. I. (2000). Individual differences in problem solving via insight. *Current Psychology*, 19, 143–146.
- Ansburg, P. I., & Dominowski, R. L. (2000). Promoting insightful problem solving. *Journal of Creative Behavior*, 34, 30–60.
- Ansburg, P. I., & Hill, K. (2003). Creative and analytic thinkers differ in their use of attentional resources. *Personality and Individual Differences*, 34, 1141–1152.
- Ash, I. K., & Wiley, J. (2006). The nature of restructuring in insight: An individual-differences approach. *Psychonomic Bulletin and Review*, 13, 66–73.
- Atchley, R. A., Keeney, M., & Burgess, C. (1999). Cerebral hemispheric mechanisms linking ambiguous word meaning retrieval and creativity. *Brain and Cognition*, 40, 479–499.
- Aziz-Zadeh, L., Kaplan, J. T., & Iacoboni, M. (2009). 'Aha!': The neural correlates of verbal insight solutions. *Human Brain Mapping*, 30, 908–916.
- Beeman, M. J., & Bowden, E. M. (2000). The hemisphere maintains solution-related activation for yet-to-be-solved problems. *Memory and Cognition*, 28, 1231–1241.
- Bowden, E. M., & Beeman, M. J. (1998). Getting the right idea: RH contributions to solving insight problems. *Psychological Science*, 9, 435–440.
- Bowden, E. M., & Jung-Beeman, M. (2003). Aha! Insight experience correlates with solution activation in the right hemisphere. *Psychonomic Bulletin and Review*, 10, 730–737.
- Carter, C. S., Macdonald, A. M., Botvinick, M., Ross, L. L., Stenger, V. A., Noll, D., & Cohen, J. D. (2000). Parsing executive processes: Strategic vs. evaluative functions of the anterior cingulate cortex. *Proceedings of the National Academy of Sciences USA*, 97, 1944–1948.

- Chen, Z., & Daehler, M. W. (2000). External and internal instantiation of abstract information facilitates transfer in insight problem solving. *Contemporary Educational Psychology*, 25, 423–449.
- Christensen, B. T., & Schunn, C. D. (2005). Spontaneous access and analogical incubation effects. *Creativity Research Journal*, 17, 207–220.
- Chronicle, E. P., MacGregor, J. N., & Ormerod, T. C. (2004). What makes an insight problem? The roles of heuristics, goal conception, and solution recoding in knowledge-lean problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 14–27.
- Chronicle, E. P., Ormerod, T. C., & MacGregor, J. N. (2001). When insight just won't come: The failure of visual cues in the nine-dot problem. *Quarterly Journal of Experimental Psychology*, 54A, 903–919.
- Clapham, M. M. (2001). The effects of affect manipulation and information exposure on divergent thinking. *Creativity Research Journal*, 13, 335–350.
- Cunningham, J. B., & MacGregor, J. N. (2008). Training insightful problem solving: Effects of realistic and puzzle-like contexts. *Creativity Research Journal*, 20, 291–296.
- Davidson, J. E. (1995). The suddenness of insight. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight* (pp. 125–155). Cambridge, MA: MIT Press.
- DeYoung, C. G., Flanders, J. L., & Peterson, J. B. (2008). Cognitive abilities involved in insight problem solving: An individual differences model. *Creativity Research Journal*, 20, 278–290.
- Dominowski, R. L., & Dallob, P. I. (1995). Insight and problem solving. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight* (pp. 33–62). Cambridge, MA: MIT Press.
- Dow, G. T., & Mayer, R. E. (2004). Teaching students to solve insight problems: Evidence for domain specificity in creativity training. *Creativity Research Journal*, 16, 389–402.
- Duncker, K. (1945). On problem-solving. *Psychological Monographs*, 58(5): Whole No. 270.
- Durso, F. T., Rea, C. B., & Dayton, T. (1994). Graph-theoretic confirmation of restructuring during insight. *Psychological Science*, 5, 94–98.
- Fiore, S. M., & Schooler, J. W. (1998). Right hemisphere contributions to creative problem solving: Converging evidence for divergent thinking. In M. Beeman & C. Chiarello (Eds.), *Right hemisphere language comprehension: Perspectives from cognitive neuroscience* (pp. 349–371). Mahwah, NJ: Erlbaum.
- Fleck, J. I. (2008). Working memory demands in insight versus analytic problem solving. *European Journal of Cognitive Psychology*, 20, 139–176.
- Fleck, J. I., & Weisberg, R. W. (2004). The use of verbal protocols as data: An analysis of insight in the candle problem. *Memory and Cognition*, 32, 990–1006.
- Friedman, R. S., & Förster, J. (2005). Effects of motivational cues on perceptual asymmetry: Implications for creativity and analytical problem solving. *Journal of Personality and Social Psychology*, 88, 263–275.
- Gaeta, H., Friedman, D., & Hunt, G. (2003). Stimulus characteristics and task category dissociate the anterior and posterior aspects of novelty P3. *Psychophysiology*, 40, 198–208.
- Gazzaniga, M. S. (1998). The split brain revisited. *Scientific American*, 279, 51–55.
- Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, 12, 306–355.
- Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, 15, 1–38.
- Gilhooly, K. J., Fioratou, E., & Henretty, N. (2010). Verbalization and problem solving: Insight and spatial factors. *British Journal of Psychology*, 101, 81–93.
- Gilhooly, K. J., & Murphy, P. (2005). Differentiating insight from noninsight problems. *Thinking and Reasoning*, 11, 279–302.
- Grant, E. R., & Spivey, M. J. (2003). Eye movements and problem solving: Guiding attention guides thought. *Psychological Science*, 14, 462–466.
- Guilford, J. P. (1950). Creativity. *American Psychologist*, 5, 444–454.
- Holyoak, K. J. (2005). Analogy. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 117–142). New York: Cambridge University Press.
- Howard-Jones, P. A., Blakemore, S. J., Samuel, E. A., Summers, I. R., & Claxton, G. (2005). Semantic divergence and creative story generation: An fMRI investigation. *Cognitive Brain Research*, 25, 240–250.
- Isen, A. M., Daubman, K. A., & Nowicki, G. P. (1987). PA facilitates creative problem solving. *Journal of Personality and Social Psychology*, 52, 1112–1131.
- Jones, G. (2003). Testing two cognitive theories of insight. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 1017–1027.
- Jung-Beeman, M. (2005). Bilateral brain processes for comprehending natural language. *Trends in Cognitive Sciences*, 9, 512–518.
- Jung-Beeman, M., Bowden, E. M., Haberman, J., Frymiare, J. L., Arambel-Liu, S., Greenblatt, R., ... Kounios, J. (2004). Neural activity when people solve verbal problems with insight. *PLoS Biology*, 2(4), e97. doi:10.1371/journal.pbio.0020097.
- Kaplan, C. A., & Simon, H. A. (1990). In search of insight. *Cognitive Psychology*, 22, 374–419.
- Karimi, Z., Windmann, S., Güntürkün, O., & Abraham, A. (2007). Insight problem solving in individuals with high versus low schizotypy. *Journal of Research in Personality*, 41, 473–480.
- Keane, M. T. (1987). On retrieving analogues when solving problems. *Quarterly Journal of Experimental Psychology*, 39A, 29–41.
- Keefe, J. A., & Magaro, P. A. (1980). Creativity and schizophrenia: An equivalence of cognitive processing. *Journal of Abnormal Psychology*, 89, 390–398.
- Kershaw, T. C., & Ohlsson, S. (2004). Multiple causes of difficulty in insight: The case of the nine-dot problem. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 3–13.
- Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 1534–1555.
- Knoblich, G., Ohlsson, S., & Raney, G. E. (2001). An eye movement study of insight problem solving. *Memory and Cognition*, 29, 1000–1009.
- Koffka, K. (1935). *Principles of Gestalt psychology*. New York: Harcourt, Brace, & Co.
- Köhler, W. (1925). *The mentality of apes*. London: Routledge and Kegan Paul.

- Kohn, N., & Smith, S. M. (2009). Partly versus completely out of your mind: Effects of incubation and distraction on resolving fixation. *Journal of Creative Behavior*, 43, 102–118.
- Kounios, J., Fleck, J. I., Green, D. L., Payne, L., Stevenson, J. L., Bowden, E., & Jung-Beeman, M. (2008). The origins of insight in resting-state brain activity. *Neuropsychologia*, 46, 281–291.
- Kounios, J., Frymiare, J. L., Bowden, E. M., Fleck, J. I., Subramaniam, K., Parrish, T. B., & Jung-Beeman, M. (2006). The prepared mind: Neural activity prior to problem presentation predicts subsequent solution by sudden insight. *Psychological Science*, 17, 882–890.
- Kurtz, K. J., & Loewenstein, J. (2007). Converging on a new role for analogy in problem solving and retrieval: When two problems are better than one. *Memory and Cognition*, 35, 334–341.
- Lang, S., Kanngieser, N., Jaśkowski, P., Haider, H., Rose, M., & Verleger, R. (2006). Precursors of insight in event-related brain potentials. *Journal of Cognitive Neuroscience*, 18, 2152–2166.
- Luchins, A. S. (1942). Mechanization in problem solving – The effect of Einstellung. *Psychological Monographs*, 54, 1–95.
- Lung, C. T., & Dominowski, R. L. (1985). Effects of strategy instructions and practice on nine-dot problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11, 804–811.
- Luo, J., & Knoblich, G. (2007). Studying insight problem solving with neuroscientific methods. *Methods*, 42, 77–86.
- Luo, J., Niki, K., & Phillips, S. (2004a). Neural correlates of the ‘Aha! Reaction.’ *NeuroReport*, 15, 2013–2017.
- Luo, J., Niki, K., & Phillips, S. (2004b). The function of the anterior cingulate cortex (ACC) in insightful problem solving: ACC activated less when the structure of the puzzle is known. *Journal of Chinese Societies*, 5, 195–213.
- MacGregor, J. N., Ormerod, T. C., & Chronicle, E. P. (2001). Information-processing and insight: A process model of performance on the nine-dot problem. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 176–201.
- Mai, X., Luo, J., Wu, J., & Luo, Y. (2004). “Aha!” effects in a guessing riddle task: An event-related potential study. *Human Brain Mapping*, 22, 261–270.
- Maier, N. R. F. (1930). Reasoning in humans I: On direction. *Journal of Comparative Psychology*, 10, 115–143.
- Maier, N. R. F. (1931). Reasoning in humans II: The solution of a problem and its appearance in consciousness. *Journal of Comparative Psychology*, 12, 181–194.
- Maier, N. R. F. (1933). An aspect of human reasoning. *British Journal of Psychology*, 24, 144–155.
- Mandler, J. M., & Mandler, G. (1964). *Thinking: From association to Gestalt*. New York: John Wiley & Sons.
- Markman, K. D., Lindberg, M. J., Kray, L. J., & Galinsky, A. D. (2007). Implications of counterfactual structure for creative generation and analytic problem solving. *Personality and Social Psychology Bulletin*, 33, 312–324.
- Martindale, C. (1995). Creativity and connectionism. In S. M. Smith, T. B. Ward, & R. A. Finke (Eds.), *The creative cognition approach* (pp. 249–268). Cambridge, MA: MIT Press.
- Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review*, 69, 220–232.
- Metcalfe, J., & Wiebe, D. (1987). Intuition in insight and non-insight problem solving. *Memory and Cognition*, 15, 238–246.
- Mohr, C., Graves, R. E., Gianotti, L. R. R., Pizzagalli, D., & Brugge, P. (2001). Loose but normal: A semantic association study. *Journal of Psycholinguistic Research*, 30, 475–483.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
- Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. In M. T. Keane & K. J. Gilhooly (Eds.), *Advances in the psychology of thinking* (Vol. 1, pp. 1–44). London: Harvester-Wheatsheaf.
- Ormerod, T. C., Chronicle, E. P., & MacGregor, J. N. (2006). Asymmetrical analogical transfer in insight problem solving. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, Mahwah, NJ: Lawrence Erlbaum Associates (pp. 1899–1904).
- Ormerod, T. C., MacGregor, J. N., & Chronicle, E. P. (2002). Dynamics and constraints in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 791–799.
- Payne, L., & Kounios, J. (2009). Coherent oscillatory networks supporting short-term memory retention. *Brain Research*, 1247, 126–132.
- Perkins, D. N. (1981). *The mind's best work*. Cambridge, MA: Harvard University Press.
- Qiu, J., Li, H., Yang, D., Luo, Y., Li, Y., Wu, Z., & Zhang, Q. (2008). The neural basis of insight problem solving: An event-related potential study. *Brain and Cognition*, 68, 100–106.
- Ray, W. J., & Cole, H. W. (1985). EEG alpha activity reflects attentional demands, and beta activity reflects emotional and cognitive processes. *Science*, 228, 750–752.
- Rowe, G., Hirsch, J. B., & Anderson, A. K. (2007). Positive affect increases the breadth of selective attention. *Proceedings of the National Academy of Sciences USA*, 104, 383–388.
- Runco, M. A., & Okuda, S. M. (1991). The instrumental enhancement of the flexibility and originality scores of divergent thinking tests. *Applied Cognitive Psychology*, 5, 435–441.
- St. George, M., Kutas, M., Martinez, A., & Sereno, M. I. (1999). Semantic integration in reading: Engagement of the right hemisphere during discourse processing. *Brain*, 122, 1317–1325.
- Scheerer, M. (1963). Problem solving. *Scientific American*, 208, 118–128.
- Schooler, J. W., & Melcher, J. (1995). The ineffability of insight. In S. M. Smith, T. B. Ward, & R. A. Finke (Eds.), *The creative cognition approach* (pp. 97–143). Cambridge, MA: MIT Press.
- Schooler, J. W., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General*, 122, 166–183.
- Schwert, P. M. (2007). Using sentence and picture clues to solve verbal insight problems. *Creativity Research Journal*, 19, 293–306.
- Segal, E. (2004). Incubation in insight problem solving. *Creativity Research Journal*, 16, 141–148.
- Seifert, C. M., Meyer, D. E., Davidson, N., Patalano, A. L., & Yaniv, I. (1995). Demystification of cognitive insight: Opportunistic assimilation and the prepared-mind perspective. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight* (pp. 157–196). Cambridge, MA: MIT Press.
- Sio, U. N., & Ormerod, T. C. (2009). Does incubation enhance problem solving? A meta-analytic review. *Psychological Review*, 135, 94–120.
- Smith, S. M. (1995). Getting into and out of mental ruts: A theory of fixation, incubation, and insight. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight* (pp. 229–251). Cambridge, MA: MIT Press.
- Smith, S. M., & Blankenship, S. E. (1991). Incubation and the persistence of fixation in problem solving. *American Journal of Psychology*, 104, 61–87.

- Smith, R.W., & Kounios, J. (1996). Sudden insight: All-or-none processing revealed by speed-accuracy decomposition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 1443–1462.
- Sternberg, R. J., & Davidson, J. E. (1982). The mind of the puzzler. *Psychology Today*, 16, 37–44.
- Subramaniam, K., Kounios, J., Parrish, T. B., & Jung-Beeman, M. (2009). A brain mechanism for facilitation of insight by positive affect. *Journal of Cognitive Neuroscience*, 21, 415–432.
- Thomas, L. E., & Lleras, A. (2007). Moving eyes and moving thought: On the spatial compatibility between eye movements and cognition. *Psychonomic Bulletin and Review*, 14, 663–668.
- Thomas, L. E., & Lleras, A. (2009a). Covert shifts of attention function as an implicit aid to insight. *Cognition*, 111, 168–174.
- Thomas, L. E., & Lleras, A. (2009b). Swinging into thought: Directed movement guides insight in problem solving. *Psychonomic Bulletin and Review*, 16, 719–723.
- Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K., & Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: A reevaluation. *Proceedings of the National Academy of Sciences USA*, 94, 1492–1497.
- Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Review: Series of Monograph Supplements*, 2(4), Whole No. 8.
- van Steenburgh, J. J. (2011). *Direct current stimulation of right anterior superior temporal gyrus during solution of compound remote associates problems*. Unpublished doctoral dissertation, Drexel University.
- Vartanian, O., & Goel, V. (2005). Task constraints modulate activation in right ventral lateral prefrontal cortex. *NeuroImage*, 27, 927–933.
- Vos, S. H., Gunter, T. C., Kohk, H. H. J., & Mulder, G. (2001). Working memory constraints on syntactic processing: An electrophysiological investigation. *Psychophysiology*, 38, 41–63.
- Vul, E., & Pashler, H. (2007). Incubation benefits only after people have been misdirected. *Memory and Cognition*, 35, 701–710.
- Wallas, G. (1926). *The art of thought*. London: Cape.
- Wertheimer, M. (1945/1959). *Productive thinking* (Enlarged ed.). London: Tavistock Publications.
- Weisberg, R. W. (1986). *Creativity: Genius and other myths*. New York: Freeman.
- Weisberg, R. W. (1995). Prolegomena to theories of insight in problem solving: A taxonomy of problems. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight* (pp. 157–196). Cambridge, MA: MIT Press.
- Weisberg, R. W. (2006). *Creativity: Understanding innovation in problem solving, science, invention, and the arts*. Hoboken, NJ: Wiley.
- Weisberg, R. W., & Alba, J. W. (1981). An examination of the alleged role of “fixation” in the solution of several “insight” problems. *Journal of Experimental Psychology: General*, 110, 169–192.
- Weisberg, R. W., & Suls, J. M. (1973). An information-processing model of Duncker's candle problem. *Cognitive Psychology*, 4, 255–276.
- Wu, L., Knoblich, G., Wei, G., & Luo, J. (2009). How perceptual processes help to generate new meaning: An EEG study of chunk decomposition in Chinese characters. *Brain Research*, 1296, 104–112.
- Yaniv, I., & Meyer, D. E. (1987). Activation and metacognition of inaccessible stored information: Potential bases for incubation effects in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 187–205.

Dean Keith Simonton

Abstract

Scientific research on genius began in the early 19th century and increased in popularity throughout the end of the century and the beginning of the 20th century. Although the first investigations used mainly historiometric methods, later psychologists introduced psychometric and experimental techniques. Definitions of genius fall into two categories: superlative intellect and phenomenal achievement, where the latter can be subdivided into extraordinary creativity, exceptional leadership, and prodigious performance. However defined, genius has been studied from four main psychological perspectives: general intelligence, domain expertise, heuristic search, and blind variation. Each of these perspectives has distinct advantages and disadvantages as explanatory accounts. As a consequence, a comprehensive understanding of how geniuses think and reason will require an integration of all four perspectives. The chapter closes with a discussion of future directions for research.

Key Words: genius, intelligence, IQ, creativity, leadership, prodigies

Introduction

As a psychological phenomenon, genius has been around for centuries. The first genius to leave a name in the historical records was perhaps Imhotep, an Egyptian polymath who lived in the 27th century BCE. The pharaoh's vizier, Imhotep attained fame as the first physician, engineer, and architect—most famous for building the step pyramid of Djoser. Of course, over the centuries the inventory of celebrated geniuses has become very large. Every world civilization produces its own list of prime candidates (Kroeber, 1944; Murray, 2003). In European civilization, names like Isaac Newton, René Descartes, William Shakespeare, Michelangelo Buonarroti, and Ludwig van Beethoven come immediately to mind. Indeed, one could argue that any given civilization is largely defined by the works that its geniuses have created. How much would our perceptions of European civilization change without the *Principia Mathematica* (the Three Laws of Motion),

the *Discourse on Method* ("I think, therefore I am"), *Hamlet* ("To be or not to be"), the Sistine Chapel frescoes (The Creation of Adam), or the Fifth Symphony (the duh-duh-duh-dah "fate knocks at the door" theme)? Such contributions have become iconic.

Although the scientific study of genius is much more recent, the first empirical investigation dates back to the early 19th century, making it one of the oldest research topics in the psychological sciences. Research began when the Belgian mathematician Adolphe Quetelet (1835/1968) published a quantitative study of the careers of eminent English and French playwrights. However, because this investigation was tucked away in a much larger work on "social statistics," Quetelet's contribution to the subject is most often overlooked. The first book-length study entirely and explicitly devoted to genius is Francis Galton's (1869) classic *Hereditary Genius*, an expansion of a journal article he had published

4 years earlier (Galton, 1865).¹ By the end of the 19th century and the beginning of the 20th century the topic of genius had attracted the attention of several notable scientists, some of whom have their own claim to genius (e.g., Cattell, 1903; Ellis, 1904; Freud, 1910/1964; Lombroso, 1891).

Many of these early investigations used historiometric methods. That is, the investigators subjected biographical and historical data to objective and quantitative analyses (Simonton, 1999c). In the 1920s, with the introduction of psychometric assessment, scientific research on genius could pursue a different direction. In 1916, Lewis Terman had devised the Stanford-Binet intelligence scale, and 5 years later he used the measure to identify more than 1,500 high-IQ boys and girls, who where then followed into adulthood. The psychometric results were reported in four volumes of *Genetic Studies of Genius* (Terman, 1925–1959; a fifth volume was a historiometric study published by Cox, 1926). Terman's classic longitudinal investigation had a profound impact on how genius was conceived both within psychological science and among the public at large. To offer one noteworthy instance, Terman's purely methodological decision to use IQ 140 as the threshold criterion for “genius” ended up in the dictionary, a point that we will return to shortly.

Finally, I should mention a third major approach to studying genius. Rather than use historiometric or psychometric methods, a few investigators have employed laboratory experiments. For example, Chase and Simon (1973) conducted an experimental investigation on short-term memory for chess positions using Hans Berliner, a former World Correspondence Chess Champion, as one of the research participants. Similarly, Jensen (1990) published an experimental study of reaction times using Shakuntala Devi, an incredible calculating prodigy who could do the arithmetic faster than she could deliver the answers. However, because contemporary geniuses are seldom willing to expose themselves to such inconvenient scrutiny, laboratory experiments are much less common than either historiometric investigations (where the raw data are archival) or psychometric studies (where the measurements can often be conducted via mail or the Internet). In fact, whether these laboratory participants can be considered “geniuses” depends on how genius is defined. It is to that issue that we turn first. After genius is defined, we will then examine the four principal perspectives on how geniuses think and reason.

Definitions

The word *genius* has a long history of etymological development (Murray, 1989). The word harks back to Roman mythology, where a genius was a person's “guardian spirit” and hence had something to do with that individual's fate and uniqueness.² Over the years, its meaning changed in multiple ways so that most dictionaries give it a half-dozen different definitions. Nonetheless, two definitions have received the most attention in psychological research, namely, phenomenal achievement and superlative intellect. Although distinct, these two usages share a common attribute: Both definitions are highly exclusive, even elitist. Although anybody in Roman times could have a genius, only a very tiny percentage of the population today can be considered to exhibit genuine genius.

Phenomenal Achievement

When Galton (1869) published the first scientific monograph on genius, his focus was clearly on those individuals who made a name for themselves for their phenomenal achievements. In particular, Galton defined *genius* in terms of achieved eminence or reputation, which he conceived as “the opinion of contemporaries, revised by posterity” that prove that given person is “a leader of opinion, . . . an originator, . . . a man [or woman] to whom the world deliberately acknowledges itself largely indebted” (p. 37). It comes as no surprise, then, that Galton listed not only outstanding creators but also exceptional leaders. The creators included eminent scientists, poets, painters, and composers, whereas the leaders included distinguished politicians, commanders, jurists, and religious figures. What is more surprising was a third group: illustrious wrestlers and oarsmen—persons who made a mark in sports! These athletes would presumably have been persons who, had they lived a century later, might have become medalists in the Olympics. Just as interesting, the chapter that appears just before those on wrestlers and oarsmen names winners of “Senior Classics” at Cambridge University, an honor comparable to “Senior Wrangler” in mathematics at the same institution. Both honors were highly competitive—essentially an intellectual Olympics (viz. the “Tripos”). If Galton were conducting the study today, he might have also included the victors of chess championships, prestigious piano competitions, and perhaps even spelling bees!

Galton's (1869) broad conception of genius compels us to consider three manifestations of

phenomenal achievement: outstanding creativity, exceptional leadership, and prodigious performance (cf. Simonton, 2009b). As will become apparent in this chapter, the three manifestations do not all have the same status as candidates for genius, nor do the three necessarily involve the same thinking and reasoning processes.

OUTSTANDING CREATIVITY

A creative idea or behavior is usually defined as having the joint properties of (a) novelty or originality and (b) utility or usefulness (Simonton, 2000b). The first attribute distinguishes creativity from the routine or mundane, the second from the eccentric or insane. Both properties are quantitative rather than qualitative attributes. For example, an idea might be novel or useful with respect to a single person, the person's family, his or her nation, or the world in general. Undoubtedly, for creativity to attain the level of genius, both novelty and utility must be truly outstanding. Not only did no one else come up with the same idea but also the utility of the idea is universal, or nearly so. Einstein's general theory of relativity is a prime example. Apropos of this instance, Immanuel Kant (1790/1952) defined the work of genius as something both original and *exemplary*. Einstein's theory is not only original but also exemplary—a paradigm that provides guidance for subsequent theoretical physicists. Obviously, Kant's definition is closely related to that of Galton.

Nevertheless, Kant (1790/1952) differed from Galton (1869) in arguing that creative genius is found in the arts but not in the sciences. Kant's exception stems from his belief that scientific creativity was governed by a method that could be learned and applied, whereas artistic creativity lacked such guidance. Artistic creators had to devise their own rules for creating exemplary art. Regardless of whether we accept Kant's distinction, it should be evident that there is no such thing as an all-inclusive psychological description of outstanding creativity (Simonton, 2009c). The thought processes by which Mendeleev devised the Periodic Table of the Elements were certainly not isomorphic with those by which his compatriot and contemporary Tchaikovsky wrote his fantasy-overture *Romeo and Juliet*. Even so, the difference may be a matter of degree rather than kind.

Although creative genius is often applied as a qualitative attribute—as if some people are geniuses and others are not—it should be emphasized that outstanding creativity can be treated as a quantitative

variable that admits degrees. For instance, magnitude of genius can be gauged by creative productivity, such as the number of high-impact works (Albert, 1975). The low end of the scale would then be anchored by those “one-hit wonders” who only manage to make a single notable contribution to their field (see, e.g., Kozbelt, 2008b). On the other hand, the highest-order geniuses would be highly prolific. The quantitative rather than qualitative status of genius also holds for the remaining two manifestations of phenomenal achievement.

EXCEPTIONAL LEADERSHIP

The concept of genius was first applied to outstanding creators, and only much later was it extended to exceptional leaders (Murray, 1989). For example, the Thomas Carlyle's (1841) inventory of “heroes” (a.k.a. “geniuses”) includes not only the creators Dante Alighieri, William Shakespeare, Jean-Jacques Rousseau, and Robert Burns but also the leaders Napoleon Bonaparte, Frederick the Great of Prussia, Oliver Cromwell, Martin Luther, and Mohammed, the Prophet of Islam. In some respects, this extension of the concept was unfortunate. For instance, exceptional leadership is far more contingent on nonpsychological situational factors—political, cultural, ideological, and so on—than holds for outstanding creativity (Simonton, 1995). As a result, variables such as general intelligence have a more ambiguous role in the case of exceptional leaders. Although there is no evidence that a creator can be too intelligent to be creative, there is reason for believing that leaders can be too intelligent to lead (Simonton, 1985, 2006). This contrast leads to another: the difference between contemporary and posthumous recognition. According to Galton (1869), genius is judged by “the opinion of contemporaries, revised by posterity” (p. 37). For leaders, both of these assessments are required, whereas only posterity's evaluation is needed for creators. Creators like poet Emily Dickinson and biologist Gregor Mendel can now bask in posthumous recognition and thereby become belatedly identified as geniuses. In contrast, those generals who had no opportunity to lead their armies in decisive battles or those heads of state who never had the chance to make history-making decisions will be forever left out of consideration. Hence, leader genius is more separated from psychological variables than is creator genius. This divergence is one reason why this form of genius will receive less attention than the others.

PRODIGIOUS PERFORMANCE

I have already suggested that Galton (1869) could have included other varieties of prodigious performers besides just celebrated athletes and prize-winning scholars. Anyone who wins competition in a highly valued domain, and by this means has his or her name set down in the corresponding “record book,” might qualify as a genius according to this more inclusive definition. This third definition might be extended to encompass those individuals who never attained such adulthood accomplishments but still established some reputation as child prodigies (Radford, 1990). A notorious illustration is William James Sidis, a mathematics wiz who never realized his potential as an adult (Montour, 1977). A counterfactual example would have been Wolfgang Amadeus Mozart had he died in his early teens, before creating his first masterworks (Hayes, 1989).

Also problematic are savants who exhibit a prodigious level of performance in some very narrow domain despite having almost no capacity to lead anything close to a normal life (Miller, 1999). This group is particularly tricky because savants, unlike most exceptional leaders and unlike all extraordinary creators, are seldom capable of producing ideas or behaviors that feature both novelty and utility. In this way, savants can also be contrasted with prodigious performers, such as a basketball point guard who devises a new tactic to circumvent a novel defense or a virtuoso violinist who conceives an ingenious phrasing for a particularly difficult cadenza. That said, the difference between a child prodigy and a savant may very well be quantitative rather than qualitative. Furthermore, some of the same cognitive processes underlying one might also underlie the other, such as extreme domain-specific memory. Where some music savants can play a composition perfectly after only a single hearing, Mozart famously wrote down the extremely complex score to Allegri’s *Miserere* with only minor corrections after hearing it only once.

Although prodigious performance probably has many more divergent manifestations than either extraordinary creativity or exceptional leadership, it is more similar to creativity in having a largely psychological foundation.

Superlative Intellect

If prodigious performance is going to be taken as a guise of genius, then what about someone whose achievement places him or her in the *Guinness Book of Records*? To be sure, some of the categories

appear far removed from anything truly worthy of the record books, such as the most paperclips strung together in a 24-hour period (Simonton, 2009b). Yet what about a more impressive accomplishment, such as having the highest recorded score on an IQ test? The latter feat would seem legitimate certification for genius status. After all, one respected dictionary followed Terman (1925–1959) by explicitly defining genius as “A person who has an exceptionally high intelligence quotient, typically above 140” (*American Heritage Electronic Dictionary*, 1992). That Marilyn Vos Savant received a Guinness-record entry for having an IQ of 228 would automatically qualify her not only as a genius but also as the world’s greatest genius (McFarlan, 1989). Because IQ tests are among the most reliable instruments in the psychometric arsenal (Stanovich, Chapter 22), defining genius as superlative intellect would seem the optimal strategy from the standpoint of psychological science. This definition has the additional asset that researchers have a much better idea about what is being assessed relative to any of the three criteria of phenomenal performance. Intelligence tests contain specific subscales that make it patent exactly what is being measured—verbal comprehension, vocabulary, general information, analogical reasoning, working memory, perceptual organization, spatial reasoning, processing speed, and so forth (depending on the specific composition of the IQ test). In addition, because all intelligence measures are designed, validated, and standardized prior to administration, it would seem reasonable to assume that these instruments provide a direct evaluation of the thinking and reasoning capacities responsible for phenomenal performance. This assumption was the actual basis for Terman’s (1925–1959) longitudinal study. After assessing his young research participants on the Stanford-Binet, he hoped to show that more than three decades later they would grow up to become adult geniuses as defined by phenomenal achievement.

We will get back to this definition shortly. But before we do, let us at least acknowledge that intelligence tests do not exhaust all the possible psychometric instruments that might be germane to the assessment of genius. Given the manifest connection between genius and outstanding creativity (Smith & Ward, Chapter 23), it is sensible to ask: What about using CQ rather than IQ? That is, why not use a creativity test to gauge genius rather than use an intelligence test? It turns out that the answer is too simple. No consensus yet exists on the optimal

psychometric scale for assessing creativity. This is not to say that no such measures have been created. On the contrary, the number of such instruments has proliferated with abandon (Simonton, 2003a). Although these assessments can exhibit modestly positive intercorrelations (e.g., Carson, Peterson, & Higgins, 2005; Gough, 1979), the relations do not approach the magnitude of alternative IQ tests (cf. McNemar, 1964). Thus, if the goal is to define genius in terms of superlative intellect, an IQ assessment at present surpasses a CQ assessment. Alternative IQ measures are far more likely to triangulate on the same candidate, whereas alternative CQ measures will point to multiple competing candidates.

Be that as it may, I must note a peculiar manner in which the superlative intellect definition is diametrically opposed to the original concept of genius. As noted earlier, genius first indicated something unique about the individual, a meaning preserved in the outstanding creativity definition, and to a lesser extent in the exceptional leadership definition. Geniuses generate something unique that sets them well apart from the crowd, whether it is Whitman's *Leaves of Grass* or Lincoln's "Gettysburg Address." Even the prodigious performance criterion leaves open the option that someone might accomplish something exceptional in a unique and even exemplary manner—such as Dick Fosbury introducing the "Fosbury flop" that completely revolutionized the high jump. Yet the use of an IQ test does not allow for such individualism. Although there are many ways of obtaining a mediocre score on such a test (depending on which specific responses were right or wrong), there is only one way of obtaining a perfect score like Vos Savant did to earn her record IQ. More generally, the higher a person's score on a standardized test, the more that person is reasoning and thinking according to some socially established norm of optimal cognitive performance. This norm is shaped by the group on which the test was standardized and by the psychometrician who created the test in the first place. In this sense, achieving a top score on an IQ test is analogous to ringing the bell at the carnival's fair strength test. The winning ring can only be achieved one way: a perfect strike with maximum force.

Perspectives

Given the definitions discussed earlier, the next task is to identify the cognitive processes or mechanisms that provide the foundation for genius. These can be roughly collated into the following four

categories: general intelligence, domain expertise, heuristic search, and blind variation. Next I review each perspective, laying out the explanatory advantages and disadvantages of each.³ To avoid misunderstanding at the outset, these categories are not mutually exclusive. Quite the opposite, genius will often entail some mix of two or more (see, e.g., Kozbelt, 2008a). The specific nature of the mixture will depend heavily on the particular type of genius. Perhaps only in the case of creative genius are all four perspectives highly relevant.

General Intelligence

Samuel Johnson (1781), the author of the first English dictionary, claimed, "the true Genius is a mind of large general powers, accidentally determined to some particular direction" (p. 5). Expressed in modern terms, geniuses are those who would score high on a test of general intelligence. They then apply this generic cognitive capacity to an arbitrarily selected domain, and phenomenal achievements result. Actually, this concept is not far removed from that held by Galton (1869), who assumed that geniuses were those who enjoyed unusually high "natural ability," so high that they were in the upper tail of the population distribution. Galton argued that it was almost inevitable that an individual with that level of ability would go on to achieve great things. Such intellects just needed to discover the proper subject on which to apply their brilliance. That choice very likely would depend on childhood and adolescent interests and hobbies (e.g., Roe, 1953; Schaefer & Anastasi, 1968) as well as early exposure to role models and mentors (e.g., Bloom, 1985; Simonton, 1984a; Walberg, Rasher, & Parkerson, 1980).

ADVANTAGES

First, the most obvious asset of this explanation is its simplicity—a one-size-fits-all account. Indeed, it provides the most elegant solution by effectively equating the superlative intellect and phenomenal achievement definitions. Second, the explanation also has some empirical support: Psychometric and historiometric studies show that geniuses in most domains are extremely intelligent, most having IQs high enough to fit the dictionary definition of genius (e.g., Roe, 1953; Simonton & Song, 2009). Some studies even find that the degree of achieved eminence is positively associated with the magnitude of estimated intelligence (Simonton, 2006, 2008a; Simonton & Song, 2009; Walberg, Rasher, & Hase,

1978). Third, the explanation also fits nicely with the position that intelligence represents a single latent variable most often referred to as “Spearman’s *g*” (cf. Spearman, 1927), a factor that provides a broad predictor of success by a wide range of criteria (Gottfredson, 1997; Ones, Viswesvaran, & Dilchert, 2005). It would seem natural, therefore, to have the same factor apply to achieved eminence. Fourth, because we supposedly know what goes into a test of general intelligence, it should be relatively easy to infer the cognitive processes that provide the basis for phenomenal achievement (e.g., Jensen, 1992). Presumably, those test subscales that load highest on *g* would assess the most relevant cognitive capacities (e.g., the abilities that underlie performance on Ravens Advanced Progressive Matrices; see Carpenter, Just, & Shell, 1990). Fifth and last, to the extent that genius depends on general intelligence, we obtain a precise answer to the classic question of whether genius is born or made—the issue that Galton (1869) first attempted to answer. The answer is directly expressed by *g*’s heritability coefficient, as estimated in behavior genetics (e.g., Bouchard & McGue, 1981). By this criterion, genius is very much (even if not entirely) born.

DISADVANTAGES

In the first place, not every psychologist subscribes to a single-factor theory of intelligence; rather, many view intelligence as having multiple factors (e.g., Guilford, 1967; Sternberg, 1996). In the specific context of genius studies, the most visible dissenter is Howard Gardner (1983, 1993), whose theory of multiple intelligences allots a distinctive intellect to Albert Einstein (logico-mathematical), T. S. Eliot (linguistic), Pablo Picasso (visual), Igor Stravinsky (musical), Martha Graham (bodily-kinesthetic), Sigmund Freud (intrapersonal), and Mahatma Gandhi (interpersonal). Gardner’s theory thus provides a separate intelligence for exceptional leadership (interpersonal) and even implies an intelligence appropriate for great athletes (bodily-kinesthetic). Gardner (1998) has also speculated about the existence of additional intelligences beyond these seven, namely, naturalist, spiritual, and existential intelligences. If Gardner is correct, then different geniuses do not think and reason in the same manner. Their thoughts have distinct modalities and patterns.

Secondly, it is manifest that genius as phenomenal achievement cannot be taken as equivalent to genius as superlative intellect, at least not as defined by an IQ test. Exceptions exist at both ends of the

intelligence distribution. At one extreme is Marilyn Vos Savant who, though boasting the word’s highest IQ, has yet to produce anything that would ensure an enduring reputation in the absence of her record-breaking IQ score. At the other extreme is someone like William Shockley who did not obtain an IQ score high enough to enter Terman’s (1925–1959) sample and yet still managed to earn a Nobel Prize for Physics as co-inventor of the transistor (Eysenck, 1995). Nobody who was admitted into Terman’s sample attained so exalted an honor. Just to show that Shockley’s case is not unique, a sizable proportion of phenomenal achievers do not have IQs in the genius range (e.g., Cox, 1926; Roe, 1953; Simonton, 2006). A nontrivial minority would not even be smart enough to join Mensa, the high IQ society that stipulates a threshold of around IQ 130 (depending on the test). In sum, the IQ definition yields both false positives and false negatives.

One reason why general cognitive ability corresponds so imperfectly with actual high-grade accomplishments is that the concept of intelligence probably must be broadened to encompass dispositional variables. Galton (1869) himself defined natural ability to include not just intellect but also energy and determination. Cox’s (1926) historiometric study of 301 geniuses concluded that drive, persistence, and determination—or what has more recently been termed GRIT (Duckworth, Peterson, Matthews, & Kelly, 2007)—can readily compensate for a person lacking a stellar IQ. Others have widened the concept of intelligence to include the “intelligent personality” (Chamorro-Premuzic & Furnham, 2006). A critical portion of this intelligent personality must entail openness to experience, a factor that may be more predictive of phenomenal achievement than is intelligence alone (Cassandra & Simonton, 2010; Rubenzer, Faschingbauer, & Ones, 2000; Simonton, 2006; see also McCrae, 1987). Closely related to openness is conceptual or integrative complexity, a cognitive style variable shown to have some predictive value for both extraordinary creators and exceptional leaders (Feist, 1994; Simonton, 2006; Suedfeld, 1985; Suedfeld, Corteen, & McCormick, 1986). Someone high on complexity can view the world from multiple perspectives and then integrate them into a comprehensive perspective—something like the elusive goal of the current chapter!

Aside from all of these considerations, we have to recognize that phenomenal achievement is almost invariably dependent on first attaining sufficient

domain-specific expertise. In all likelihood, this is biggest single reason why Vos Savant, for all her superlative intellect, has been unable to produce a phenomenal achievement. She is a Jane of all trades, master of none.

Domain Expertise

A very large empirical and theoretical literature has developed on expert performance and the acquisition of domain-specific expertise (Ericsson, 1996; Ericsson, Charness, Feltovich, & Hoffman, 2006). With respect to this chapter, the bulk of the work appears most applicable to instances of prodigious performance, such as becoming a virtuoso pianist or a chess grandmaster (e.g., Chase & Simon, 1973; de Groot, 1978; Ericsson, Krampe, & Tesch-Römer, 1993). Even so, the core ideas can easily be extrapolated to the two other forms of phenomenal achievement, especially outstanding creativity. The basic argument is that (a) phenomenal achievement first requires the acquisition of domain-specific knowledge and skills and (b) this domain-specific expertise necessitates approximately a full decade of deliberate practice and study—the “10-year” rule. Only after persons get beyond this period of extensive apprenticeship can they be expected to have the expertise needed to make world-class contributions to their chosen domains. To illustrate, Hayes (1989) studied 76 famous classical composers and determined the age when they began the intensive study of music and the age when they contributed their first unquestioned masterpiece to the repertoire. Typically, these two developmental landmarks were separated by 10 years (see also Simonton, 1991b). This rule even applied to a precocious composer like Mozart.⁴

ADVANTAGES

First and foremost, by viewing genius as dependent on domain-specific expertise, this perspective overthrows a commonplace myth about genius, namely that it is utterly born rather than made—a position dating back to Galton (1869). Instead, genius in any domain certainly requires hard work in the form of study and practice. This requirement is especially crucial for outstanding creativity (where assiduous study is more critical) and prodigious performance (where dedicated practice may be more imperative), but it is probably important in many domains of extraordinary leadership as well. It is certainly no accident that the general most likely to emerge victorious on the battlefield is the leader with the most experience leading his or her forces

in combat (Simonton, 1980). This is not to say that luck or serendipity never plays a role. Some persons are fortunate to be at the right place at the right time (Simonton, 1994). But such individuals seldom attain the highest levels in the annals of genius. As an example, Karl Jansky, an engineer with Bell Telephone Laboratories, made a chance observation that directly contributed to the development of radio astronomy. Yet he is not considered a major figure in astronomy (Murray, 2003). He had no expertise in the domain.

A second significant advantage of this approach is that it provides a strong basis for empirical research. Investigators can examine the specific components of a given domain expertise as well as the optimal procedures for acquiring that expertise (Ericsson, Charness, Feltovich, & Hoffman, 2006). For example, Chase and Simon (1973) conducted a now classic study of chess expertise that showed that while an expert player was better able to reconstruct chess positions when they represented actual games, an expert did no better than novices when the pieces were placed randomly on the board (see also de Groot, 1978). Experts can quickly “chunk” familiar but not unfamiliar arrangements. This acquired chunking ability also provided the basis for Shakuntala Devi’s prodigious calculating abilities (Jensen, 1990). She evidently cannot even view a multiple-digit number without instantaneously breaking it down into factors! Yet it took her a great many years to acquire this capacity. She was not born with the talent.

So impressive are such findings that they have led some researchers to believe that genius can be totally explicated in terms of acquired domain-specific expertise (e.g., Howe, 1999). In a sense, “innate talent” does not exist (e.g., Howe, Davidson, & Sloboda, 1998). If Einstein is considered one of the greatest physicists of all time, it is because he studied harder and longer than his contemporaries. The world’s best student must necessarily become the world’s most praised genius. Valedictorians always come out victorious.

DISADVANTAGES

The first problem with the domain expertise perspective is implied by the close of the preceding paragraph: Its advocates too often take their position to the extreme, arguing for the sole impact of “nurture” with as much misguided enthusiasm as Galton (1869) had argued for “nature.” Although empirical research has shown that eminence is often

enhanced by promoting extreme positions, including on the nature-nurture issue (Simonton, 1976b, 2000c), that is probably not the optimal route to acquiring a full understanding of such a complex phenomenon. Rather than view them as mutually exclusive, it is probably more productive to integrate the two into a single coherent perspective. For instance, it is possible to define talent in terms of expertise acquisition (Simonton, 2008b). Someone can be considered talented if he or she inherited a set of cognitive abilities and dispositional traits that allow him or her to (a) accelerate the acquisition of domain-specific expertise or (b) attain higher performance levels with a given amount of domain-specific expertise. As this definition has it, talent cannot even be defined without assuming the acquisition of domain expertise. It otherwise has no meaning.⁵

In line with the preceding problem, there are certain empirical characteristics of genius that cannot be accommodated entirely within the domain expertise perspective (Simonton, 2000a). This explanatory deficiency is especially conspicuous in the case of extraordinary creativity. One should infer that the optimal strategy would be extreme specialization, concentrating expertise on a specialty area so as to become the world's leading expert in that domain. Yet the opposite is the case: Creative geniuses tend to be much more versatile than the norm, having both extremely wide interests and the exceptional capacity to contribute to multiple domains or subdomains (e.g., Cassandro, 1998; Root-Bernstein, Bernstein, & Garnier, 1993, 1995; Root-Bernstein et al., 2008; Simonton, 1976a, 1976b, 2000a; Sulloway, 1996; White, 1931). The same versatility connection seems to hold for exceptional leadership as well (Simonton, 1976a, 2006; White, 1931). If the 10-year rule applied in these cases, then they would need two or more lifetimes to accomplish what they achieved in a single career. This inconsistency with straightforward domain expertise implies that breadth of knowledge and openness to experience contribute to creative thinking in a manner that has no parallel in those domains that rely exclusively on deliberate practice. This inference is reinforced by what was said earlier about the intelligent personality. For certain kinds of genius, something more is required than simply "practice makes perfect."

Heuristic Search

To get around the limitations of the foregoing two perspectives, we can call up the notion that

some forms of genius are engaged in problem solving of a very high order (e.g., Weisberg, 1992, 2006). This enables us to draw upon the theoretical and empirical work on problem solving pioneered by Newell, Simon, and their colleagues and students (e.g., Klahr & Simon, 1999; Newell, Shaw, & Simon, 1958; Newell & Simon, 1972). Put in a nutshell, researchers in the Newell-Simon tradition make an instructive distinction between "strong" and "weak" methods in problem solving (Bassok & Novick, Chapter 21).

At one extreme, strong methods tend to be highly domain specific but also highly algorithmic, so that the methods pretty much guarantee a solution if competently applied. To offer a simple illustration, anyone with a sufficient amount of algebra can provide the solution ("roots" or "zeros") for the equation $y = 3x^2 - 2x + 10$. The numbers 3, -2, and 10 only have to be plugged into the appropriate spots in the quadratic formula.

At the other extreme, unusually difficult problems—problems that are far more novel and complex—do not easily lend themselves to such tried and true techniques. On the contrary, problem solvers will then have to rely on weak or "heuristic" methods. Unlike algorithmic methods, heuristic methods are often more generic in the sense that they are often applicable to a wide range of problems. Yet partly because of that very domain generality, none of them can guarantee a solution, and it may not always be self-evident which heuristic method is applicable to a given problem. Heuristic methods of the broadest scope include generate and test (or trial and error), hill climbing (or steepest ascent), means-end analysis, and analogy (Klahr & Simon, 1999). Analogy, for instance, has been shown to play a major role in the thinking of scientific geniuses (Gentner & Jeziorski, 1989; Dunbar & Klahr, Chapter 35). A light "wave" is an analogy (Holyoak, Chapter 13).

Needless to say, if creativity is defined by novelty and utility, then heuristic methods are far more likely to produce creative solutions than are algorithmic methods. Algorithms can guarantee a solution, but they cannot assure that the solution will be original. Indeed, although the generate-and-test strategy is probably the weakest of all methods, it frequently assumes primary importance. As Newell, Shaw, and Simon (1962) said, "In spite of the primitive character of trial-and-error processes, they bulk very large in highly creative problem-solving; in fact, at the upper end of the range of problem difficulty

there is likely to be a positive correlation between creativity and the use of trial-and-error generators” (pp. 72–73). A portion of this process will entail trying out alternative heuristic methods and thereby elevate generate and test to what has been called a “trial-and-error meta-heuristic” (Simonton, 2004).

ADVANTAGES

Perhaps the most prominent asset of this perspective is that it demystifies the concept of creative genius (Hayes, 1989). Because creativity is nothing more than ordinary problem solving “writ large,” researchers do not have to call upon any mysterious processes like some “stroke of genius” or “flash of inspiration.” The contrast between everyday creativity and genius-level creativity is quantitative rather than qualitative. Marie Curie did not inherently think any differently than the average person on the street but rather merely applied a set of strong and weak methods to a particular area of expertise. As this statement implies, the heuristic search perspective also connects nicely with the domain expertise perspective, as would be expected given that Herbert Simon directly contributed to both viewpoints. The two perspectives together can explain all three forms of phenomenal achievement, exceptional leadership and prodigious performance as well as extraordinary creativity. These three guises of genius would simply involve a different mix of domain-specific strong methods along with a differential dependence on weak methods. For example, the chess champion probably depends more on game-specific strong methods, whereas the artistic genius probably leans more on the application of more generic weak methods, including trial and error (Simonton, 2007). This contrast may be partially responsible for a striking contrast: Although computers have yet to write poetry deserving of a Nobel Prize or even Pulitzer Prize, computers have beaten the best human chess player in the world (Hsu, 2002). Among other reasons, chess genius is much more algorithmic than poetic genius is.

Another major advantage of the heuristic search perspective is its capacity for inspiring research. Much of this research entails laboratory experiments on problem solving (e.g., Qin & Simon, 1990), but a considerable literature has also been devoted to developing and testing computational models (e.g., Langley, Simon, Bradshaw, & Zythew, 1987; Shrager & Langley, 1990). Of special interest are the discovery programs that purport to replicate the actual creative achievements of geniuses, many

of the programs being explicitly named after those geniuses, such as OCCAM, BACON, GALILEO, GLAUBER, STAHL, FAHRENHEIT, BLACK, and DALTON. A case in point is BACON, a program that has made numerous data-driven or inductive discoveries (Bradshaw, Langley, & Simon, 1983). Of these achievements, the program’s rediscovery of Kepler’s Third Law of Planetary Motion has probably received the most attention. This law relates the planet’s period of rotation around the sun, P , to its distance from the sun, D , according to $P^2 = kD^3$, where k is a proportionality constant. When given the raw data for the planets, BACON was able to arrive at the same law. Furthermore, rather than use a brute-force approach that would examine all possible mathematical functions, the program conducted a heuristic search that restricted the problem space (Simon, 1983). Given that Kepler scores as one of the greatest geniuses in the history of astronomy (Murray, 2003), and given that his three laws are considered among his most significant contributions, BACON can be said to have provided a computer simulation of scientific genius.

It should be noted that the computer program BACON has no domain-specific expertise, so this simulation would seem to contradict the domain expertise perspective. Because BACON also does not display any general intelligence whatsoever, the simulation challenges that perspective as well. Thus, one final advantage of the heuristic search perspective is its seeming egalitarianism: Almost anybody can become a genius. Neither a stratospheric IQ nor 10 years of domain expertise is required. Ironically, although Herbert Simon can certainly be considered to have been one of the great geniuses in the history of cognitive science—as polymath and Nobel laureate—he put much emphasis on this point. For instance, Simon (1973) held that “Mendeleev’s Periodic Table does not involve a notion of pattern more complex than that required to handle patterned letter sequences” (p. 479). Similarly, Simon (1986) conducted an informal experiment indicating that nothing out of the ordinary was needed to discover something that would win the discovery’s originator a Nobel Prize:

On eight occasions I have sat down at lunch with colleagues who are good applied mathematicians and said to them: “I have a problem that you can perhaps help me with. I have some very nice data that can be fitted very accurately for large values of the independent variable by an exponential function, but

for small values they fit a linear function accurately. Can you suggest a smooth function that will give me a good fit through the whole range?" (p. 7)

Of the eight lunch companions, five got a solution in only a few minutes. Not one wondered what Simon was surreptitiously about, nor did any recognize the historic status of the problem they were given. Even so, those five colleagues had independently derived Max Planck's equation for black body radiation. In a separate clandestine experiment, a graduate student in chemical engineering derived the Balmer formula for the hydrogen spectrum (Qin & Simon, 1990). Genius does not amount to much if historic achievements can be replicated so effortlessly by anonymous faculty and graduate students. Although Simon does not provide enough information to discern whether any of his participants had genius-level IQs, he does provide sufficient reason to conclude that none of them had any domain-specific expertise.

DISADVANTAGES

As should be apparent by now, one problem with the heuristic search perspective is that in the process of demystifying genius it ends up trivializing genius. It is not simple to score at stratospheric levels on an IQ test nor is it easy to put in the 10 years of hard labor necessary to attain first-class expertise in an achievement domain, and yet BACON was able to derive in a few seconds what it took Kepler many years to discover. However, this derivation ignores the far richer biographical and historical context in which this discovery was actually made. The Third Law emerged out of a complex matrix of ideas (including the first two laws and his unique physical interpretation of the heliocentric system), and the law was tested on observational data inferior to those fed into BACON. What appears obvious in retrospect was far from obvious in Kepler's day. Indeed, Kepler's astronomy was not fully appreciated by his older contemporary Galileo (who insisted on circular rather than elliptical orbits). It is not even clear that BACON's heuristics would have been available to Kepler in the early 17th century. Kepler may have had to discover from scratch what BACON had programmed a priori into its operation. If this trivialization of genius were justified, then genius would be far more common (rather than less common) than implied by the superior intellect definition. As a rough estimate, rather than represent 1% or less of the population, geniuses should constitute 10% or more.

A second problem concerns this perspective's limited focus. Unlike the domain expertise perspective, research on heuristic search has concentrated mostly on creative problem solving, and often more specifically on scientific discovery (Klahr & Simon, 1999). The two other forms of phenomenal achievement—outstanding leadership and prodigious performance—receive much less attention, if any. Yet it is fair to say that this limitation is not that critical. Insofar as genius-grade leadership requires superior problem-solving skills, there is no reason why the heuristic search perspective cannot provide valuable insights. And, with regard to prodigious performance, it is probably likely that problem solving per se is much less prominent, and when problems do present themselves they might be best solved through strong methods.

A final difficulty with the heuristic search perspective is the insufficient attention it gives to individual-difference variables. Although contrasts in general intelligence might be acknowledged as well as variation in acquired expertise, personality factors are largely ignored, if not outright denied (e.g., Klahr & Simon, 1999; for further discussion, see Simonton, 2003b). Yet as pointed out earlier, phenomenal achievement is associated with a large number of dispositional variables. Although some of these variables may not be particularly germane to understanding how geniuses think and reason (e.g., motivation; but see Amabile, 1996), others are certainly relevant (e.g., cognitive style). We will return to this matter shortly after we present the fourth and final perspective.

Blind Variation

Donald Campbell (1960) proposed that creative thought could be explicated in terms of the two-step process of blind variation and selective retention, or BVSR (cf. BV+SR in Nickles, 2003). In essence, the creator generates two or more "thought trials" or "ideational variants" without secure knowledge of their probable outcomes, and then selects those that appear most useful. The connection between BVSR and the definition of creativity should be obvious: The BV step generates the idea's novelty and the SR step determines the idea's utility. It is essential to acknowledge the following two characteristics of Campbell's original BVSR theory. First, blind variation does not require that the ideational process be random. Although all random processes are necessarily blind, even systematic searches can be blind as well (e.g., radar scans and search grids). Second,

the theory is not predicated on an analogy with biological evolution. On the contrary, Campbell claims that the original prototype of BVSR creativity may be found in Bain (1855/1977), who published his theory 4 years before Darwin's *Origin of Species*. Hence, it is incorrect to call BVSR "Darwinian" (cf. Kronfeldner, 2010; Simonton, 1999b). Indeed, it might be more correct to call it "Popperian" given its close connection with Karl Popper's philosophy (Dennett, 1995; e.g., Popper, 1972).

Unfortunately, Campbell (1960) never fully developed BVSR as a theory of creativity, but instead turned to questions of sociocultural evolution and evolutionary epistemology (e.g., Campbell, 1965, 1974). However, over the past couple of decades Simonton has published a series of target articles (1999a, 2007, 2010), book chapters (e.g., 1993, 2009a), and books (1988, 1999b) that gradually converted BVSR into a comprehensive theory of creative genius in both the arts and sciences. So it is Simonton's version of Campbell's theory that we here refer to as the blind variation perspective on genius. One asset of the updated version is that Simonton (2010, 2011a) provided a formal (mathematical) definition of "blindness" that removed a regrettable imprecision in Campbell's informal (verbal) definition. That definitional vagueness had inspired too many unnecessary objections to the BVSR position (e.g., Kronfeldner, 2010; Sternberg, 1998; Thagard, 1988).

Like the previous three perspectives, the blind variation perspective has both advantages and disadvantages.

ADVANTAGES

To begin with, the blind variation perspective is integrative: Rather than restrict creativity to a single process, the theory recruits a wide range of processes to generate the ideational variations, the only stipulation being that they have some degree of blindness (Simonton, 1999b, 2011b). According to this perspective, the thinking and reasoning processes underlying genius-level creativity can be placed along a continuous dimension from total sightedness to complete blindness (Simonton, 2011a). In the parlance of the heuristic search perspective, problem-solving methods can range from algorithmic strong methods to an exceedingly weak method like generate and test, with the heuristic procedures used by BACON to discover Kepler's Third Law falling in the middle (Simonton, 2011b). Other processes or procedures far more likely to reside near the blind end of this dimension include (a) remote

associations (Mednick, 1962), (b) rare associations (Gough, 1976), (c) divergent thinking of various kinds (e.g., "unusual uses;" Guilford, 1967), (d) primordial or primary process cognition (Martindale, 1990; Suler, 1980), (e) Janusian associations and homospatial imagery (Rothenberg, 1979), (f) allusive or overinclusive thought (Eysenck, 1995), (g) clang associations (Hadamard, 1945), and (h) behavioral tinkering (Kantorovich, 1993). An illustration of the last blind-variation procedure is how James Watson came up with the DNA code via the trial-and-error manipulation of cardboard models for the four bases (Simonton, 2011b). In line with what Newell, Shaw, and Simon (1962) suggested earlier, the more extraordinary the creativity, the higher is the likelihood that the participating processes operated closer to the blind end of the spectrum. Accordingly, the blind variation perspective purports to be especially useful in comprehending creative genius of the highest order.

Besides being integrative, the blind variation perspective claims considerable explanatory power (Simonton, 1999b, 2010). That is, many aspects about creativity and creative genius become more interpretable within a BVSR framework. For instance, the perspective is most consistent with several individual-difference variables associated with creative achievement, such as why creativity is positively associated with reduced latent inhibition (e.g., Carson, Peterson, & Higgins, 2003). Highly creative people often come up with ideas in states of defocused attention in which they notice things they are supposed to ignore (e.g., how a certain mold ruins bacteria cultures). Because reduced latent inhibition is also associated with psychoticism and schizotypy (Eysenck, 1995), this connection provides a basis for comprehending why exceptional creativity is often linked to some degree of psychopathological symptoms, especially among artistic geniuses and revolutionary scientists (Ko & Kim, 2008; Ludwig, 1992; Post, 1994). Perhaps for this reason highly creative individuals are more prone to come from family lineages with higher than average rates of psychopathology (e.g., Carlson, 1970; Richards, Kinney, Lunde, Benet, & Merzel, 1988). Creativity and psychopathology have been shown to share a genetic basis (Kéri, 2009; see also Eysenck, 1995).

In addition, the blind variation perspective accounts for certain key developmental factors (Simonton, 1999b, 2010). As an illustration, Campbell (1960) speculated that "persons who have been uprooted from traditional cultures, or who have

been thoroughly exposed to two or more cultures, seem to have the advantage in the range of hypotheses they are apt to consider, and through this means, in the frequency of creative innovation" (p. 391). This speculation enjoys empirical support (e.g., Leung, Maddux, Galinsky, & Chiu, 2008). Similarly, data are in line with Kuhn's (1970) conjecture that revolutionary scientists are often "very new to the field whose paradigm they change" (p. 90; see Jeppesen & Lakhani, 2010; Simonton, 1984b). Excessive expertise can hinder exceptional creativity if it prematurely truncates the range of ideational variants (Simonton, 2000a, 2011b). It may be for this cause that experiments find that creative performance is often enhanced when individuals are exposed to unpredictable, novel, or incongruous stimuli (Finke, Ward, & Smith, 1992; Proctor, 1993; Rothenberg, 1986; Sobel & Rothenberg, 1980; Wan & Chiu, 2002).

Finally, the blind variation perspective has had some predictive value (Simonton, 1999b, 2007, 2010). Of special importance are combinatorial models that can account for a wide range of phenomena related to extraordinary creativity, such as creative output across and within careers or the dramatic event where two or more scientists independently arrive at the same discovery (Simonton, 2004, 2010). It is also worth mentioning in this context the existence of computer programs that operate according to BVSR principles (Goldberg, 1989; Holland, 1975, 1992; Koza, 1992). These programs have successfully planned fiber-optic telecommunication networks, made forecasts in currency trading, designed steam and gas turbines, enhanced oil exploration and mining operations, and improved the efficiency of jet engines (Holland, 1992), and solved difficult problems in algebraic equations, the determination of animal foraging behavior, the design of electrical circuits, and the identification of optimal game-playing strategies (Koza, 1994; Koza, Bennett III, Andre, & Keane, 1999). One of these programs even rediscovered Kepler's Third Law (Koza, 1992). Thus, these blind variation programs more than hold their own relative to the discovery programs that implement heuristic searches. Given that all computer programs that simulate creativity have so far relied on some kind of random generator (Boden, 2004), it would seem that blind variation has a critical place in extraordinary creativity (Simonton, 2010, 2011b).

DISADVANTAGES

The primary drawback of the blind variation perspective is that it has only been extensively developed

for just one form of genius, namely extraordinary creativity. Although it might be at least partially applicable to exceptional leadership as well, such an application remains to be worked out. The perspective is not likely to have any value for understanding most forms of prodigious performance. Whereas the improvisations of a virtuoso jazz musician probably depend somewhat on blind variation, it is hard to imagine the value of such processes for a chess champion or Olympic athlete. Finally, it would be absurd to hold that superlative intellect was contingent on BVSR. In general, the more a domain of genius is contingent on acquired expertise (including intelligence test performance as a very special form of expertise), then the less relevant BVSR becomes as an explanatory perspective. The blind variation perspective is confined to situations when the genius must "think outside the box," "go beyond the information given," or engage in "lateral thinking." Hence, blind variation largely begins where expertise ends. Nevertheless, insofar as extraordinary creativity counts as the most prototypical form of the phenomenon of genius, this limitation may not be all that irksome.

Paradoxically, a second shortcoming of the blind variation perspective ensues from one of its strong points: its very integrative nature. BVSR theory is inclusive rather than exclusive, accommodating so many diverse forms of thinking and reasoning—explicit and implicit, voluntary and involuntary, logical and irrational, systematic and stochastic, and so on. Hence, a critic of the perspective must seriously wonder whether it has any conceptual integrity. This skepticism might only be reinforced rather than reduced by the intimate connection between the BVSR theory of creativity and "universal selection theory" (Cziko, 1995, 1998, 2001). A theory that applies with equal success to biological evolution, the immune system, neurological development, operant conditioning, and other phenomena cannot possibly provide more than a superficial account of genius, creative or otherwise. Consilience can certainly be a virtue (Wilson, 1998). Yet such knowledge unification can also be a vice. By explaining everything, a theory might just end up explaining nothing.

The last failing could be connected with a third flaw: In comparison to the previous three perspectives, the blind variation perspective has not stimulated nearly as much empirical research. Moreover, the bulk of the research regarding the model has been historiometric rather than psychometric or experimental (e.g., Simonton, 1991a, 2007). Yet it is conceivable that this deficiency might be remedied in the

future as increasingly more investigators come to view creative genius from this perspective. Alternatively, a critic could very well argue that the blind variation perspective is so inclusive that it cannot support empirically decisive tests. One critic of the blind variation perspective even said that it represents a religion rather than a science (Sternberg, 1999).

Conclusions

Genius can be defined in two principal ways: as superlative intellect and as phenomenal achievement. Furthermore, the second definition can encompass extraordinary creativity, exceptional leadership, or prodigious performance. Given this conceptual heterogeneity, it is understandable that the phenomenon has been examined from at least four major perspectives: general intelligence, domain expertise, heuristic search, and blind variation. Each perspective has variable utility depending on the definition of genius. This linkage between definition and perspective is most obvious in the case of general intelligence and superlative intellect because the latter tends to be most frequently defined by the former. The domain expertise perspective probably has the greatest explanatory value for understanding prodigious performance, whereas the heuristic search and blind variation perspectives prove most useful in discussing extraordinary creativity.

It is apparent that exceptional leadership has received somewhat less attention as a guise of genius. Because leadership is an intrinsically social rather than individual phenomenon, it is reasonable to assume that much more is involved relative to the other kinds of genius. This increased complexity could entail individual characteristics as well as the situational factors touched upon earlier. For instance, charisma no doubt has a role to play in exceptional leadership (Deluga, 1997, 1998; Emrich, Brower, Feldman, & Garland, 2001; House, Spangler, & Woycke, 1991). Yet charisma probably has a minimal place in extraordinary creativity or prodigious performance—and certainly not in superlative intellect. Nobody needs charisma to score at the highest levels on an IQ test.

Future Directions

A great deal has been learned about genius since the days of Quetelet, Galton, and Terman. Yet much more needs to be learned as well. Certainly more scientific work is needed on how geniuses think and reason. This additional research should most fruitfully pursue the following three lines of inquiry:

1. With genius coming in four varieties—superlative intellect and three types of phenomenal performance (extraordinary creativity, extraordinary leadership, and prodigious performance)—perhaps the most vital question is their empirical interconnections. To what extent are the three kinds of phenomenal performance contingent on not just superlative intellect but also various subsidiary forms of cognitive ability, such as spatial reasoning (see Hegarty & Stull, Chapter 31)? To what degree does extraordinary leadership depend on extraordinary creativity? Can charismatic leadership be conceived as prodigious performance?

2. In a like manner, with four major perspectives on genius, of whatever variety, an equally central issue is how these perspectives interrelate not only with each other but also with the four varieties of genius. For example, how does general intelligence relate with the three other perspectives? Is there a minimum IQ necessary to acquire domain-specific expertise, and does that intelligence threshold depend on the form of expertise? Do different heuristic methods have differential applicability to diverse forms of phenomenal achievement? How does blind variation intermesh with the application of heuristics? Does it enter passively or tacitly by default? Can the capacity for blind variation be acquired as a part of a person's expertise?

3. The resolution of the previous two enigmas must then be connected with several related issues. As an example, how exactly do various individual-difference variables—whether cognitive style or personality traits—impinge on the thinking and reasoning processes behind phenomenal achievement? We know that openness to experience is a consistent correlate of both creativity and leadership, but how specifically does openness affect domain expertise, heuristic search, or blind variation? Can the same individual-difference variable have a positive impact on one but a negative impact on their other? A closely related issue is the relative influence of nature and nurture. For example, to what extent is conceptual or integrative complexity, which is associated with both creativity and leadership, a function of genetic and environmental influences?

A caveat is now in order. Scientific psychology has long been split into two independent subdisciplines (Cronbach, 1957; see also Tracy, Robins, & Sherman,

2009). On the one side of the division is experimental psychology, with its interest in the general human mind—a tradition that harks back to Wilhelm Wundt, the reputed founder of experimental methods. On the other side of the division is correlational psychology, with its fascination with how human beings vary and the way those individual differences covary—a tradition dating back to Francis Galton, the founder of correlational methods. These two psychologies not only have distinct methods (e.g., analysis of variance versus regression analysis) but distinctive perspectives on psychological phenomena. These differences in outlook show up in the research on genius. For example, where the general intelligence perspective belongs to correlational psychology, the domain expertise and heuristic search perspectives belongs to experimental psychology. The blind variation perspective is the only one of the four to try to bridge the gap. Regardless of whether that specific integration was successful, it should be clear that a comprehensive understanding of genius must require researchers to aspire to such an integration of the two scientific psychologies.

Notes

1. One could argue that Liu Shao's *Jen wu chih* (*The Study of Human Abilities*), written around the mid-3rd century CE, might have historical priority, but having read both Liu and Galton, I must side with the latter. Although Liu is admirably systematic and objective, he fails to specify any core hypotheses, systematically collect any data, or subject any such data to quantification. The same contrasts apply to other candidates for writing the first scientific work solely devoted to genius as a psychological phenomenon. Galton was the first.

2. To be precise, men had a genius and women a juno, but the latter term did not contribute to the later development of the concept. There are female geniuses, but no female junii.

3. It will become apparent that the evaluation of the four perspectives will depend on neither qualitative single-case studies of geniuses nor introspective reports of those geniuses about their thought processes. Although such impressionistic information can sometimes provide useful illustrations, they cannot be considered reliable scientific data. Such data can be used for or against any perspective (e.g., Dasgupta, 2004; Simonton, 2004; Weisberg, 2006).

4. Hayes (1989) committed one conspicuous error, however, saying that "Albeniz's first masterwork was written in the 72nd year of his career!" (p. 296). That composer died before his 49th birthday.

5. The participants in the Study of Mathematically Precocious Youth (SMPY) may eventually illustrate how incredibly accelerated expertise acquisition translates into adulthood genius (Lubinski & Benbow, 1994; Lubinski, Webb, Morelock, & Benbow, 2001; Wai, Lubinski, & Benbow, 2005). Although this longitudinal inquiry still has some time to go before it reaches the scope of the Terman (1925–1959) investigation, it has already

produced one recipient of the Fields Medal, the so-called "Nobel Prize of Mathematics" (Muratori et al., 2006).

References

- Albert, R. S. (1975). Toward a behavioral definition of genius. *American Psychologist*, 30, 140–151.
- Amabile, T. M. (1996). *Creativity in context: Update to the social psychology of creativity*. Boulder, CO: Westview.
- American heritage electronic dictionary* (3rd ed.). (1992). Boston, MA: Houghton Mifflin.
- Bain, A. (1977). *The senses and the intellect* (D. N. Robinson, Ed.). Washington, DC: University Publications of America. (Original work published 1855)
- Bloom, B. S. (Ed.). (1985). *Developing talent in young people*. New York: Ballantine Books.
- Boden, M. A. (2004). *The creative mind: Myths & mechanisms* (2nd ed.). New York: Routledge.
- Bouchard, T. J., Jr., & McGue, M. (1981). Familial studies of intelligence. *Science*, 212, 1055–1059.
- Bradshaw, G. F., Langley, P. W., & Simon, H. A. (1983). Studying scientific discovery by computer simulation. *Science*, 222, 971–975.
- Campbell, D. T. (1960). Blind variation and selective retention in creative thought as in other knowledge processes. *Psychological Review*, 67, 380–400.
- Campbell, D. T. (1965). Variation and selective retention in socio-cultural evolution. In H. R. Barringer, G. I. Blanksten, & R. W. Mack (Eds.), *Social change in developing areas* (pp. 19–49). Cambridge, MA: Schenkman.
- Campbell, D. T. (1974). Evolutionary epistemology. In P. A. Schlippe (Ed.), *The philosophy of Karl Popper* (pp. 413–463). La Salle, IL: Open Court.
- Carlyle, T. (1841). *On heroes, hero-worship, and the heroic*. London: Fraser.
- Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, 97, 404–431.
- Carson, S., Peterson, J. B., & Higgins, D. M. (2003). Decreased latent inhibition is associated with increased creative achievement in high-functioning individuals. *Journal of Personality and Social Psychology*, 85, 499–506.
- Carson, S., Peterson, J. B., & Higgins, D. M. (2005). Reliability, validity, and factor structure of the Creative Achievement Questionnaire. *Creativity Research Journal*, 17, 37–50.
- Cassandra, V. J. (1998). Explaining premature mortality across fields of creative endeavor. *Journal of Personality*, 66, 805–833.
- Cassandra, V. J., & Simonton, D. K. (2010). Versatility, openness to experience, and topical diversity in creative products: An exploratory historiometric analysis of scientists, philosophers, and writers. *Journal of Creative Behavior*, 44, 1–18.
- Cattell, J. M. (1903). A statistical study of eminent men. *Popular Science Monthly*, 62, 359–377.
- Chamorro-Premuzic, T., & Furnham, A. (2006). Intellectual competence and intelligent personality: A third way in differential psychology. *Review of General Psychology*, 10, 251–267.
- Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, 4, 55–81.
- Cox, C. (1926). *The early mental traits of three hundred geniuses*. Stanford, CA: Stanford University Press.

- Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, 12, 671–684.
- Cziko, G. A. (1995). *Without miracles: Universal selection theory and the second Darwinian revolution*. Cambridge, MA: MIT Press.
- Cziko, G. A. (1998). From blind to creative: In defense of Donald Campbell's selectionist theory of human creativity. *Journal of Creative Behavior*, 32, 192–208.
- Cziko, G. A. (2001). Universal selection theory and the complementarity of different types of blind variation and selective retention. In C. Heyes, & D. L. Hull (Eds.), *Selection theory and social construction: The evolutionary naturalistic epistemology of Donald T. Campbell* (pp. 15–34). Albany: State University of New York Press.
- Dasgupta, S. (2004). Is creativity a Darwinian process? *Creativity Research Journal*, 16, 403–413.
- de Groot, A. D. (1978). *Thought and choice in chess* (2nd ed.). The Hague, Netherlands: Mouton.
- Deluga, R. J. (1997). Relationship among American presidential charismatic leadership, narcissism, and related performance. *Leadership Quarterly*, 8, 51–65.
- Deluga, R. J. (1998). American presidential proactivity, charismatic leadership, and rated performance. *Leadership Quarterly*, 9, 265–291.
- Dennett, D. C. (1995). *Darwin's dangerous idea: Evolution and the meanings of life*. New York: Simon & Schuster.
- Emrich, C. G., Brower, H. H., Feldman, J. M., & Garland, H. (2001). Images in words: Presidential rhetoric, charisma, and greatness. *Administrative Science Quarterly*, 46, 527–557.
- Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). GRIT: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, 92, 1087–1101.
- Ellis, H. (1904). *A study of British genius*. London: Hurst & Blackett.
- Ericsson, K. A. (Ed.). (1996). *The road to expert performance: Empirical evidence from the arts and sciences, sports, and games*. Mahwah, NJ: Erlbaum.
- Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (Eds.). (2006). *The Cambridge handbook of expertise and expert performance*. New York: Cambridge University Press.
- Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100, 363–406.
- Eysenck, H. J. (1995). *Genius: The natural history of creativity*. Cambridge, England: Cambridge University Press.
- Feist, G. J. (1994). Personality and working style predictors of integrative complexity: A study of scientists' thinking about research and teaching. *Journal of Personality and Social Psychology*, 67, 474–484.
- Finke, R. A., Ward, T. B., & Smith, S. M. (1992). *Creative cognition: Theory, research, applications*. Cambridge, MA: MIT Press.
- Freud, S. (1964). *Leonardo da Vinci and a memory of his childhood* (A. Tyson, Trans.). New York: Norton. (Original work published 1910.)
- Galton, F. (1865). Hereditary talent and character. *Macmillan's Magazine*, 12, 157–166, 318–327.
- Galton, F. (1869). *Hereditary genius: An inquiry into its laws and consequences*. London: Macmillan.
- Gardner, H. (1983). *Frames of mind: A theory of multiple intelligences*. New York: Basic Books.
- Gardner, H. (1993). *Creating minds: An anatomy of creativity seen through the lives of Freud, Einstein, Picasso, Stravinsky, Eliot, Graham, and Gandhi*. New York: Basic Books.
- Gardner, H. (1998). Are there additional intelligences? The case for naturalist, spiritual, and existential intelligences. In J. Kane (Ed.), *Education, information, and transformation* (pp. 111–131). Upper Saddle River, NJ: Merrill.
- Gentner, D., & Jeziorski, M. (1989). Historical shifts in the use of analogy in science. In B. Ghoshal, W. R. Shadish, Jr., R. A. Neimeyer, & A. C. Houts (Eds.), *The psychology of science: Contributions to metascience* (pp. 296–325). Cambridge, England: Cambridge University Press.
- Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning*. Reading, MA: Addison-Wesley.
- Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence*, 24, 79–132.
- Gough, H. G. (1976). Studying creativity by means of word association tests. *Journal of Applied Psychology*, 61, 348–353.
- Gough, H. G. (1979). A creative personality scale for the adjective check list. *Journal of Personality and Social Psychology*, 37, 1398–1405.
- Guilford, J. P. (1967). *The nature of human intelligence*. New York: McGraw-Hill.
- Hadamard, J. (1945). *The psychology of invention in the mathematical field*. Princeton, NJ: Princeton University Press.
- Hayes, J. R. (1989). *The complete problem solver* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Holland, J. (1975). *Natural and artificial systems*. Ann Arbor, MI: University of Michigan Press.
- Holland, J. H. (1992). Genetic algorithms. *Scientific American*, 267(1), 66–72.
- House, R. J., Spangler, W. D., & Woycke, J. (1991). Personality and charisma in the U.S. presidency: A psychological theory of leader effectiveness. *Administrative Science Quarterly*, 36, 364–396.
- Howe, M. J. A. (1999). *Genius explained*. Cambridge, England: Cambridge University Press.
- Howe, M. J. A., Davidson, J. W., & Sloboda, J. A. (1998). Innate talents: Reality or myth? *Behavioral and Brain Sciences*, 21, 399–442.
- Hsu, F. (2002). *Behind Deep Blue: Building the computer that defeated the world chess champion*. Princeton, NJ: Princeton University Press.
- Jensen, A. R. (1990). Speed of information processing in a calculating prodigy. *Intelligence*, 14, 259–274.
- Jensen, A. R. (1992). Understanding g in terms of information processing. *Educational Psychology Review*, 4, 271–308.
- Jeppesen, L. B., & Lakhani, K. R. (2010). Marginality and problem-solving effectiveness in broadcast search. *Organization Science*, 21, 1016–1033.
- Johnson, S. (1781). *The lives of the most eminent English poets* (Vol. 1). London: Bathurst.
- Kant, I. (1952). The critique of judgement. In R. M. Hutchins (Ed.), *Great books of the Western world* (Vol. 42, pp. 459–613). Chicago, IL: Encyclopaedia Britannica. (Original work published 1790)
- Kantorovich, A. (1993). *Scientific discovery: Logic and tinkering*. Albany: State University of New York Press.
- Karlson, J. I. (1970). Genetic association of giftedness and creativity with schizophrenia. *Hereditas*, 66, 177–182.
- Kéri, S. (2009). Genes for psychosis and creativity: A promoter polymorphism of the *Neuregulin 1* gene is related to creativity in people with high intellectual achievement. *Psychological Science*, 20, 1070–1073.

- Klahr, D., & Simon, H. A. (1999). Studies of scientific creativity: Complementary approaches and convergent findings. *Psychological Bulletin, 125*, 524–543.
- Ko, Y., & Kim, J. (2008). Scientific geniuses' psychopathology as a moderator in the relation between creative contribution types and eminence. *Creativity Research Journal, 20*, 251–261.
- Koza, J. R. (1992). *Genetic programming: On the programming of computers by means of natural selection*. Cambridge, MA: MIT Press.
- Koza, J. R. (1994). *Genetic programming II: Automatic discovery of reusable programs*. Cambridge, MA: MIT Press.
- Koza, J. R., Bennett, F. H., III, Andre, D., & Keane, M. A. (1999). *Genetic programming III: Darwinian invention and problem solving*. San Francisco, CA: Morgan Kaufmann.
- Kozbelt, A. (2008a). Longitudinal hit ratios of classical composers: Reconciling "Darwinian" and expertise acquisition perspectives on lifespan creativity. *Psychology of Aesthetics, Creativity, and the Arts, 2*, 221–235.
- Kozbelt, A. (2008b). One-hit wonders in classical music: Evidence and (partial) explanations for an early career peak. *Creativity Research Journal, 20*, 179–195.
- Kroeber, A. L. (1944). *Configurations of culture growth*. Berkeley: University of California Press.
- Kronfeldner, M. E. (2010). Darwinian 'blind' hypothesis formation revisited. *Synthese, 175*, 193–218. doi: 10.1007/s11229-009-9498-8.
- Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago, IL: University of Chicago Press.
- Langley, P., Simon, H. A., Bradshaw, G. L., & Zythow, J. M. (1987). *Scientific discovery*. Cambridge, MA: MIT Press.
- Leung, A. K., Maddux, W. W., Galinsky, A. D., & Chiu, C. (2008). Multicultural experience enhances creativity: The when and how. *American Psychologist, 63*, 169–181.
- Lombroso, C. (1891). *The man of genius*. London: Scott.
- Lubinski, D., & Benbow, C. P. (1994). The study of mathematically precocious youth: The first three decades of a planned 50-year study of intellectual talent. In R. F. Subotnik & K. D. Arnold (Eds.), *Beyond Terman: Contemporary longitudinal studies of giftedness and talent* (pp. 255–281). Norwood, NJ: Ablex.
- Lubinski, D., Webb, R. M., Morelock, M. J., & Benbow, C. P. (2001). Top 1 in 10,000: A 10-year follow-up of the profoundly gifted. *Journal of Applied Psychology, 86*, 718–729.
- Ludwig, A. M. (1992). Creative achievement and psychopathology: Comparison among professions. *American Journal of Psychotherapy, 46*, 330–356.
- Martindale, C. (1990). *The clockwork muse: The predictability of artistic styles*. New York: Basic Books.
- McCrae, R. R. (1987). Creativity, divergent thinking, and openness to experience. *Journal of Personality and Social Psychology, 52*, 1258–1265.
- McFarlan, D. (Ed.). (1989). *Guinness book of world records*. New York: Bantam.
- McNemar, Q. (1964). Lost: Our intelligence? Why? *American Psychologist, 19*, 871–882.
- Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review, 69*, 220–232.
- Miller, L. K. (1999). The Savant Syndrome: Intellectual impairment and exceptional skill. *Psychological Bulletin, 125*, 31–46.
- Montour, K. (1977). William James Sidis, the broken twig. *American Psychologist, 32*, 265–279.
- Muratori, M. C., Stanley, J. C., Gross, M. U. M., Ng, L., Tao, T., Ng, J., & Tao, B. (2006). Insights from SMPY's greatest former prodigies: Drs. Terence ("Terry") Tao and Lenhard ("Lenny") Ng reflect on their talent development. *Gifted Child Quarterly, 50*, 307–324.
- Murray, C. (2003). *Human accomplishment: The pursuit of excellence in the arts and sciences, 800 B.C. to 1950*. New York: HarperCollins.
- Murray, P. (Ed.). (1989). *Genius: The history of an idea*. Oxford, England: Blackwell.
- Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review, 65*, 151–166.
- Newell, A., Shaw, J. C., & Simon, H. A. (1962). The processes of creative thinking. In H. E. Gruber, G. Terrell, & M. Wertheimer (Eds.), *Contemporary approaches to creative thinking* (pp. 63–119). New York: Atherton Press.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
- Nickles, T. (2003). Evolutionary models of innovation and the Meno problem. In L. V. Shavinina (Ed.), *The international handbook on innovation* (pp. 54–78). New York: Elsevier Science.
- Ones, D. S., Viswesvaran, C., & Dilchert, S. (2005). Cognitive ability in selection decisions. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 431–468). Thousand Oaks, CA: Sage Publications.
- Popper, K. (1972). *Objective knowledge: An evolutionary approach*. Oxford, England: Clarendon Press.
- Post, F. (1994). Creativity and psychopathology: A study of 291 world-famous men. *British Journal of Psychiatry, 165*, 22–34.
- Proctor, R. A. (1993). Computer stimulated associations. *Creativity Research Journal, 6*, 391–400.
- Qin, Y., & Simon, H. A. (1990). Laboratory replication of scientific discovery processes. *Cognitive Science, 14*, 281–312.
- Quételet, A. (1968). *A treatise on man and the development of his faculties*. New York: Franklin. (Reprint of 1842 Edinburgh translation of 1835 French original)
- Radford, J. (1990). *Child prodigies and exceptional early achievers*. New York: Basic Books.
- Richards, R., Kinney, D. K., Lunde, I., Benet, M., & Merzel, A. P. C. (1988). Creativity in manic-depressives, cyclothymes, their normal relatives, and control subjects. *Journal of Abnormal Psychology, 97*, 281–288.
- Roe, A. (1953). *The making of a scientist*. New York: Dodd, Mead.
- Root-Bernstein, R., Allen, L., Beach, L., Bhadula, R., Fast, J., Hosey, C.,... Weinlander, S. (2008). Arts foster scientific success: Avocations of Nobel, National Academy, Royal Society, and Sigma Xi members. *Journal of the Psychology of Science and Technology, 1*, 51–63.
- Root-Bernstein, R. S., Bernstein, M., & Garnier, H. (1993). Identification of scientists making long-term, high-impact contributions, with notes on their methods of working. *Creativity Research Journal, 6*, 329–343.
- Root-Bernstein, R. S., Bernstein, M., & Garnier, H. (1995). Correlations between avocations, scientific style, work habits, and professional impact of scientists. *Creativity Research Journal, 8*, 115–137.
- Rothenberg, A. (1979). *The emerging goddess: The creative process in art, science, and other fields*. Chicago, IL: University of Chicago Press.

- Rothenberg, A. (1986). Artistic creation as stimulated by superimposed versus combined-composite visual images. *Journal of Personality and Social Psychology*, 50, 370–381.
- Rubenzer, S. J., Faschingbauer, T. R., & Ones, D. S. (2000). Assessing the U.S. presidents using the revised NEO Personality Inventory. *Assessment*, 7, 403–420.
- Schaefer, C. E., & Anastasi, A. (1968). A biographical inventory for identifying creativity in adolescent boys. *Journal of Applied Psychology*, 58, 42–48.
- Shrager, J., & Langley, P. (Eds.). (1990). *Computational models of scientific discovery and theory formation*. San Mateo, CA: Kaufmann.
- Simon, H. A. (1973). Does scientific discovery have a logic? *Philosophy of Science*, 40, 471–480.
- Simon, H. A. (1983). Discovery, invention, and development: Human creative thinking. *Proceedings of the National Academy of Sciences USA*, 80, 4569–4571.
- Simon, H. A. (1986). What we know about the creative process. In R. L. Kuhn (Ed.), *Frontiers in creative and innovative management* (pp. 3–20). Cambridge, MA: Ballinger.
- Simonton, D. K. (1976a). Biographical determinants of achieved eminence: A multivariate approach to the Cox data. *Journal of Personality and Social Psychology*, 33, 218–226.
- Simonton, D. K. (1976b). Philosophical eminence, beliefs, and zeitgeist: An individual-generational analysis. *Journal of Personality and Social Psychology*, 34, 630–640.
- Simonton, D. K. (1980). Land battles, generals, and armies: Individual and situational determinants of victory and casualties. *Journal of Personality and Social Psychology*, 38, 110–119.
- Simonton, D. K. (1984a). Artistic creativity and interpersonal relationships across and within generations. *Journal of Personality and Social Psychology*, 46, 1273–1286.
- Simonton, D. K. (1984b). Is the marginality effect all that marginal? *Social Studies of Science*, 14, 621–622.
- Simonton, D. K. (1985). Intelligence and personal influence in groups: Four nonlinear models. *Psychological Review*, 92, 532–547.
- Simonton, D. K. (1988). *Scientific genius: A psychology of science*. Cambridge, England: Cambridge University Press.
- Simonton, D. K. (1991a). Career landmarks in science: Individual differences and interdisciplinary contrasts. *Developmental Psychology*, 27, 119–130.
- Simonton, D. K. (1991b). Emergence and realization of genius: The lives and works of 120 classical composers. *Journal of Personality and Social Psychology*, 61, 829–840.
- Simonton, D. K. (1993). Genius and chance: A Darwinian perspective. In J. Brockman (Ed.), *Creativity: The Reality Club IV* (pp. 176–201). New York: Simon & Schuster.
- Simonton, D. K. (1995). Personality and intellectual predictors of leadership. In D. H. Saklofske & M. Zeidner (Eds.), *International handbook of personality and intelligence* (pp. 739–757). New York: Plenum.
- Simonton, D. K. (1994). *Greatness: Who makes history and why*. New York: Guilford Press.
- Simonton, D. K. (1999a). Creativity as blind variation and selective retention: Is the creative process Darwinian? *Psychological Inquiry*, 10, 309–328.
- Simonton, D. K. (1999b). *Origins of genius: Darwinian perspectives on creativity*. New York: Oxford University Press.
- Simonton, D. K. (1999c). Significant samples: The psychological study of eminent individuals. *Psychological Methods*, 4, 425–451.
- Simonton, D. K. (2000a). Creative development as acquired expertise: Theoretical issues and an empirical test. *Developmental Review*, 20, 283–318.
- Simonton, D. K. (2000b). Creativity: Cognitive, developmental, personal, and social aspects. *American Psychologist*, 55, 151–158.
- Simonton, D. K. (2000c). Methodological and theoretical orientation and the long-term disciplinary impact of 54 eminent psychologists. *Review of General Psychology*, 4, 13–24.
- Simonton, D. K. (2003a). Creativity assessment. In R. Fernández-Ballesteros (Ed.), *Encyclopedia of psychological assessment* (Vol. 1, pp. 276–280). London: Sage Publications.
- Simonton, D. K. (2003b). Scientific creativity as constrained stochastic behavior: The integration of product, process, and person perspectives. *Psychological Bulletin*, 129, 475–494.
- Simonton, D. K. (2004). *Creativity in science: Chance, logic, genius, and zeitgeist*. Cambridge, England: Cambridge University Press.
- Simonton, D. K. (2006). Presidential IQ, openness, intellectual brilliance, and leadership: Estimates and correlations for 42 US chief executives. *Political Psychology*, 27, 511–639.
- Simonton, D. K. (2007). The creative imagination in Picasso's *Guernica* sketches: Monotonic improvements or nonmonotonic variants? *Creativity Research Journal*, 19, 329–344.
- Simonton, D. K. (2008a). Childhood giftedness and adulthood genius: A historiometric analysis of 291 eminent African Americans. *Gifted Child Quarterly*, 52, 243–255.
- Simonton, D. K. (2008b). Scientific talent, training, and performance: Intellect, personality, and genetic endowment. *Review of General Psychology*, 12, 28–46.
- Simonton, D. K. (2009a). Creativity as a Darwinian phenomenon: The blind-variation and selective-retention model. In M. Krausz, D. Dutton, & K. Bardsley (Eds.), *The idea of creativity* (2nd ed., pp. 63–81). Leiden, Netherlands: Brill.
- Simonton, D. K. (2009b). *Genius 101*. New York: Springer Publications.
- Simonton, D. K. (2009c). Varieties of (scientific) creativity: A hierarchical model of disposition, development, and achievement. *Perspectives on Psychological Science*, 4, 441–452.
- Simonton, D. K. (2010). Creativity as blind-variation and selective-retention: Constrained combinatorial models of exceptional creativity. *Physics of Life Reviews*, 7, 156–179.
- Simonton, D. K. (2011a). Creativity and discovery as blind variation and selective retention: Multiple-variant definitions and blind-sighted integration. *Psychology of Aesthetics, Creativity, and the Arts*, 5, 222–228.
- Simonton, D. K. (2011b). Creativity and discovery as blind variation: Campbell's (1960) BVSR model after the half-century mark. *Review of General Psychology*, 15, 158–174.
- Simonton, D. K., & Song, A. V. (2009). Eminence, IQ, physical and mental health, and achievement domain: Cox's 282 geniuses revisited. *Psychological Science*, 20, 429–434.
- Sobel, R. S., & Rothenberg, A. (1980). Artistic creation as stimulated by superimposed versus separated visual images. *Journal of Personality and Social Psychology*, 39, 953–961.
- Spearman, C. (1927). *The abilities of man: Their nature and measurement*. New York: Macmillan.
- Sternberg, R. J. (1996). *Successful intelligence*. New York: Simon & Schuster.
- Sternberg, R. J. (1998). Cognitive mechanisms in human creativity: Is variation blind or sighted? *Journal of Creative Behavior*, 32, 159–176.

- Sternberg, R. J. (1999). Darwinian creativity as a conventional religious faith. *Psychological Inquiry*, 10, 357–359.
- Suedfeld, P. (1985). APA presidential addresses: The relation of integrative complexity to historical, professional, and personal factors. *Journal of Personality and Social Psychology*, 47, 848–852.
- Suedfeld, P., Corteen, R. S., & McCormick, C. (1986). The role of integrative complexity in military leadership: Robert E. Lee and his opponents. *Journal of Applied Social Psychology*, 16, 498–507.
- Suler, J. R. (1980). Primary process thinking and creativity. *Psychological Bulletin*, 88, 144–165.
- Sulloway, F. J. (1996). *Born to rebel: Birth order, family dynamics, and creative lives*. New York: Pantheon.
- Terman, L. M. (1916). *The measurement of intelligence: An explanation of and a complete guide for the use of the Stanford revision and extension of the Binet-Simon intelligence scale*. Boston, MA: Houghton Mifflin.
- Terman, L. M. (1925–1959). *Genetic studies of genius* (5 vols.). Stanford, CA: Stanford University Press.
- Thagard, P. (1988). *Computational philosophy of science*. Cambridge, MA: MIT Press.
- Tracy, J. L., Robins, R. W., & Sherman, J. W. (2009). The practice of psychological science: Searching for Cronbach's two streams in social-personality psychology. *Journal of Personality and Social Psychology*, 96, 1206–1225.
- Wai, J., Lubinski, D., & Benbow, C. P. (2005). Creativity and occupational accomplishments among intellectually precocious youths: An age 13 to age 33 longitudinal study. *Journal of Educational Psychology*, 97, 484–492.
- Walberg, H. J., Rasher, S. P., & Hase, K. (1978). IQ correlates with high eminence. *Gifted Child Quarterly*, 22, 196–200.
- Walberg, H. J., Rasher, S. P., & Parkerson, J. (1980). Childhood and eminence. *Journal of Creative Behavior*, 13, 225–231.
- Wan, W. W. N., & Chiu, C-Y. (2002). Effects of novel conceptual combination on creativity. *Journal of Creative Behavior*, 36, 227–240.
- Weisberg, R. W. (1992). *Creativity: Beyond the myth of genius*. New York: Freeman.
- Weisberg, R. W. (2006). *Creativity: Understanding innovation in problem solving, science, invention, and the arts*. Hoboken, NJ: Wiley.
- White, R. K. (1931). The versatility of genius. *Journal of Social Psychology*, 2, 460–489.
- Wilson, E. O. (1998). *Consilience: The unity of knowledge*. New York: Alfred A. Knopf.

This page intentionally left blank

PART 5

Ontogeny, Phylogeny,
Language, and Culture

This page intentionally left blank

Development of Thinking in Children

Susan A. Gelman and Brandy N. Frazier

Abstract

Children's thinking is of broad relevance to the study of thinking and reasoning, because it provides insight into fundamental issues concerning the building blocks of cognition, the role of experience, and how conceptual change comes about. The present chapter reviews four main aspects of children's thinking: (1) developmental changes, and how best to characterize such changes; (2) early cognitive capacities in infancy and early childhood, and the methodological tools that reveal these capacities; (3) causal reasoning and naive theories in childhood; and (4) the ways in which children are prepared to learn from the actions and testimony of others. The chapter ends with open questions and directions for future research.

Key Words: children, thinking, causal reasoning, naive theories, testimony, development, Piaget

Introduction

Why Study Thinking in Children?

Many people study thinking in children because they are interested in children: They may work with children in some capacity (as a teacher, or a speech therapist, for example), or they are parents and have had opportunity to observe their own child's growth (as with Charles Darwin, who wrote detailed observations of the development of his oldest son, William), or they may simply find children to be charming and inherently interesting (recall the popularity of the television show *Kids Say the Darndest Things* in the 1990s). There are also several theoretical reasons that the study of children's thought is relevant, indeed central, to anyone interested in thinking and reasoning.

(1) *The study of children tells us how knowledge comes about and how it evolves.* The kinds of learning that take place in childhood are profound. Consider that in the first 4 years of life, children construct

mental structures to deal with number, space, time, human relationships, physical principles, and organization of the objects and events surrounding them. These achievements entail inherently developmental issues—conceptual change, critical periods, nature-nurture—that can best be studied by a focus on children.

(2) *A focus on children reveals what is intelligent in even the humblest of behaviors.* The simple act of grasping a rattle, or sucking on a hand, or retrieving a dropped toy, can reveal layers of capacity and skill (Oztop, Bradley, & Arbib, 2004; von Hofsten, 2007). This, in turn, helps sharpen the questions, What is thinking? and What are its separable components?

(3) *A focus on children tells us about human preparedness.* What is special about our species that permits us to learn certain skills rapidly and with minimal training? Language is a key example. Children are actually better language learners than adults (Johnson & Newport, 1989).

This suggests a biological preparedness to learn language. At the same time, this prepared learning takes place within certain constrained conditions (e.g., learning from live interlocutors, not televised images; Kuhl, Tsoa, & Liu, 2003).

(4) *Childhood lays bare human reasoning biases.* Human thought is limited, biased, and prone to heuristics and shortcuts (Gigerenzer, 2000; Kahneman, Slovic, & Tversky, 1982). If one focuses on adults, these limitations can sometimes be difficult to see. In children, however, they are overt. One example is that of egocentrism. Originally proposed by Piaget and Inhelder (1956) to be characteristic of an early developmental stage, it turns out to be a tendency throughout the lifetime (Keysar, 2007).

(5) *Development is a methodological tool.* The study of development answers questions that cannot otherwise be addressed when reasoning structures are well in place. Stephen Jay Gould (1983) exemplifies this point in his discussion of whether a zebra is a white animal with black stripes or a black animal with white stripes. An examination of a mature zebra provides no insight, since white and black stripes are intertwined. To answer this question, biologists have had to examine the developing zebra embryo (thereby reaching the conclusion that a zebra is a black animal with white stripes!). When focusing on human thought, developmental methods likewise permit special insights: Longitudinal methods reveal the stability of traits or beliefs over time, and microgenetic methods reveal the process of change as it unfolds (Siegler & Crowley, 1991).

Plan of This Chapter

This chapter takes a selective approach to the vast subject of the development of thinking, focusing on four main topics. First, we review the broad developmental changes in thought that take place in childhood, and we briefly discuss theoretical debates concerning how best to characterize these changes. Second, we review a sampling of the rich and varied evidence for early cognitive capacities, in infancy and early childhood. A consistent theme throughout the past 40 years has been the surprising complexity and sophistication of early thought. In this section, we also briefly touch on the importance of appropriate methodological tools for uncovering these abilities. Third, we consider the realm of causal reasoning and naive theories, as an example of the cognitive structures that children are building

as well as processes that contribute to those structures. Finally, we note the importance of social input to children's developing thought, and the ways in which children are prepared to learn from the actions and testimony of others.

Developmental Change

When interacting with children, one is immediately struck by the differences in behavior as a function of age. Just by knowing a child's age, it is possible to predict quite accurately which skills the child will and will not yet have acquired. A newborn infant can't grasp a toy or recognize herself in a mirror, let alone talk or add a series of numbers. A 5-year-old child is highly skilled at language and mirror recognition and can recall past events but typically can't yet read or multiply numbers. A 10-year-old child can read, write, multiply, and divide but doesn't seem to have much grasp of the political issues of the day, has no methodological approach to scientific problems, and reasons about moral issues in a manner that seems rather rule-bound and rigid. A 17-year-old seems on a different plane of thought, able to reason skillfully about all of the issues that elude the younger children.

The theory that is best known and most influential to describe and explain these developments is that of Jean Piaget, the prominent Swiss scholar who is widely considered the "father" of cognitive developmental psychology. Piaget is rightly credited for a number of remarkable achievements. He came up with an ambitious, overarching theory that explains all of cognitive development (as opposed to working on just one aspect or sliver of cognition), and he is still recognized as capturing fundamental differences between children of different ages. Aspects of his theory are surprisingly forward looking and validated by more recent research. Piaget envisioned children as active, constructive thinkers, not simply passively taking in information supplied by others. He emphasized the importance of intrinsic motivation (vs. reward). And, over the course of a remarkably productive career, he made an amazing variety of interesting observations about children's thinking across many topics ranging from children's understanding of physical causality to their understanding of dreams, to their understanding of morality.

It is beyond the scope of this chapter to provide a full description or evaluation of Piaget's theory (see Flavell, 1963; Ginsburg & Opper, 1969; Piaget, 2000). However, we will briefly review his theory of stages and turn to four broader questions that

underlie this framework: whether there can be said to be qualitative change over childhood in the nature of thought; whether children's thinking is domain general or domain specific; the nature of individual and cultural variation; and the effects of training and environmental experiences on cognitive development.

Piagetian Stages: A Brief Overview

Piaget proposed that children progress through four stages of thought, corresponding to distinct age periods: sensorimotor (0–2 years), preoperational (2–7 years), concrete operational (7–11 years), and formal operational (11 years +). Each of these stages is said to be characterized by certain limitations as well as certain achievements. For example, at the beginning of the sensorimotor period, infants don't realize that objects still exist when they disappear from view (i.e., they lack object permanence); at the beginning of the preoperational period, children don't realize that quantities are unchanging if their shape is transformed (i.e., they lack conservation of number, conservation of volume, etc.).

According to Piaget, stages have the following characteristics: Each stage is a qualitative advance from the previous stage; at a given stage, thinking is uniform, making use of the same structures to address a broad range of problems across a broad range of topics; the sequence of stages is unchanging and universal; and children can't be trained to move to a new stage before they're ready. Each of these assumptions has been called into question on the basis of more recent research evidence, which we review here.

Qualitative Versus Incremental Change

One tenet of stage theories is that children undergo qualitative changes in thought over time. For example, Piaget and Inhelder (1956) proposed that preoperational children are egocentric, literally having difficulty taking into account a perspective other than their own. On this view, it is not until the concrete operational stage that nonegocentric thought becomes possible. For example, when given a three-dimensional display depicting three mountains and asked to report the perspective of a person sitting across the table from themselves, preschool children typically report that the person sitting across from them will see the same spatial perspective as their own. In contrast, by 7 years of age, children will consistently report the other person's perspective correctly.

One of the central challenges to a stage theory is that Piaget underestimated children's abilities—sometimes strikingly so. Although there is controversy regarding just how capable infants and young children are (see Woodward & Needham, 2009, for discussion), there is broad agreement than children are much more capable than Piaget's initial observations would suggest (see the later section entitled "Early Competence"). This critique has been directed toward both methodological and theoretical issues. At a methodological level, the concerns are multiple. Piaget's methods were highly demanding and complex from an information-processing perspective. Children were often asked to explain their answers (thus requiring verbal and metacognitive abilities, in addition to whatever conceptual capacity was being measured). In addition, Piaget's tasks were often embedded in an unfamiliar experimental context that did not permit children to make use of more familiar strategies or contexts. (We also note, however, that novel contexts may be required to reveal reasoning as opposed to retrieval of learned routines.)

Considering once again the example of egocentrism, researchers have found that a variety of modifications can lead to significantly earlier success on a perspective-taking task, including the following: simplifying the display (e.g., presenting a single object with distinctive vantage points, such as a mouse holding a candle in one hand, rather than a scene with multiple objects; Fishbein, Lewis, & Keiffer, 1972), simplifying the task (e.g., asking the child to rotate the mouse so that the candle is facing the researcher, rather than reporting a given perspective; Fishbein et al., 1972), or embedding the task in a more natural set of actions (e.g., communicating the presence of a hidden toy to a parent who has either witnessed the hiding event or not; O'Neill, 1996).

Furthermore, careful task analysis reveals that constructs that Piaget analyzed as unitary may have different levels, so that early capacities are present much younger than Piaget would have granted (again, undermining the notion of a qualitative leap in capacity). For example, perspective-taking turns out to have at least two distinct levels that emerge at different points developmentally (Flavell, Everett, Croft, & Flavell, 1981). Level-1 perspective taking entails knowing *that* someone has a different perspective from one's own; Level-2 perspective taking entails knowing *what* that perspective is. Children may fail to figure out what perspective another person has, while still having a basically nonegocentric grasp of perspectives. Furthermore, children

18 months of age understand that different people can have competing desires (Repacholi & Gopnik, 1997), even though an understanding of competing beliefs takes longer to develop (Bartsch & Wellman, 1995; Rakoczy, Warneken, & Tomasello, 2007).

From a theoretical standpoint, researchers have criticized Piaget's assumption of developmental "dichotomies" (e.g., from concrete to abstract; Keil, Smith, Simons, & Levin, 1998). They note that the reverse developmental pattern can take place as well: An abstract conceptual "framework" may then later be filled in with specifics. For example, long before children have learned the particulars of what differentiates the insides of animals versus machines, they expect animals and machines to differ in their internal parts (Simons & Keil, 1995).

A final point of criticism is that adult thinking isn't as logical and mature as Piaget had suggested. Egocentrism provides a compelling example. Rather than being restricted to a particular developmental stage, egocentrism is found even among adults, using more subtle methods. For example, when placed in a referential communication task and given an ambiguous instruction that could be interpreted from either the speaker's perspective or from an egocentric perspective, adults first gaze toward the referent that is the egocentric interpretation (Keysar, Barr, Balin, & Brauner, 2000). Similarly, when asked to judge whether a hypothetical hiker would be in greater need of food or drink, participants implicitly take an egocentric perspective, being more likely to select "drink" when they themselves are thirsty than when they are not (Van Boven & Loewenstein, 2003).

What do these critiques of Piaget imply about the broader issue of qualitative versus incremental change? Some have concluded that there is no qualitative change in cognitive development, and that instead what is most striking is the continuity from childhood through adulthood (Keil, 1981; Spelke, 2004). Others argue for qualitative change, without the additional commitment to stages, though the nature of that change is quite varied. For example, some have proposed that children's thought changes from associative to theory based (Rakison & Lupyán, 2008; Sloutsky, 2003), whereas others argue that children's thought is more akin to theory change in science (Carey, 2009; Gopnik & Schulz, 2004). Maturational changes (e.g., in working memory; in prefrontal cortex and inhibitory control) have also been proposed to yield stage-like shifts (Case, 1992, 1995; Halford, Cowan, & Andrews, 2007; Morrison

& Knowlton, Chapter 6). These issues are among the most central puzzles of cognitive development.

Domain Generality Versus Domain Specificity

Piaget proposed that thinking at a given stage was "all of a piece"—a child who is preoperational when reasoning about quantity is also preoperational when reasoning about morality, gender, classification, or perspective taking. In other words, he assumed that *domain-general* principles characterize thought. Although domain-general principles are important, more recent evidence shows that domain-specific principles are also important. For example, children do better than adults at learning a second language, whereas adults do better than children at learning higher mathematics, thus suggesting that different processes are at work in these different domains (Gelman & Noles, 2011).

At least three sources have been offered to account for domain differences in cognitive development: modularity, experience and expertise, and causal theories. Modularity is the position that certain cognitive processes are biologically constrained as the result of our evolutionary heritage. Development in these areas is thought to be highly predictable, making use of innate structures and domain-specific processes. Often (though not necessarily) these domains are associated with particular brain regions. A classic example is that of face perception, which emerges early and predictably in infancy and is associated with the fusiform face area (FFA) in the brain (Kanwisher & Yovel, 2009). Other cognitive processes that have been hypothesized to reflect modularity are language acquisition (Pinker, 1994) or numerical reasoning (Feigenson, Dehaene, & Spelke, 2004).

Expertise can also exert powerful domain-specific effects in children's cognitive performance. A striking example is that a child chess expert outperforms an adult chess novice, when it comes to memory for chess pieces on a chessboard (Chi, 1978). These differences disappear when examining memory outside the domain of expertise (e.g., digit span). One area where expertise effects have significant long-term implications is that of becoming a skilled reader, and there is great interest in understanding the changes that take place with increasing literacy experiences (Treiman & Kessler, 2007).

A final perspective on domain specificity is that children construct causal knowledge structures, often referred to as naïve theories, that reflect the particular

ontology and causal processes of a particular domain (e.g., a theory of mind concerns thoughts and beliefs; a theory of physics concerns objects and forces). On this view, children are seen as analogous to scientists who devise theories to explain phenomena around them and revise those theories in light of contradictory evidence. However, young children differ from scientists in several important respects: They do not systematically test the hypotheses they entertain, and they do not engage in precise measurement or disconfirmation of alternative hypotheses (Kuhn et al., 1988). Nonetheless, children are like scientists in positing basic ontological distinctions and making causal predictions (Wellman & Gelman, 1998). Furthermore, children take in new evidence and revise their hypotheses on the basis of such evidence, as can be seen in the layers of increasing detail and specificity in infants' understanding of physical support (Baillargeon, Li, Ng, & Yuan, 2009).

Individual and Cultural Variation

One area that has not received much research attention concerns the nature of individual differences in cognitive development, as well as cultural variation in the processes of developmental change. Of particular interest is whether these simply involve changes in *rate* of development (e.g., some individuals achieving a certain milestone earlier than others), or whether there are qualitative changes in the *process or outcome* of development.

Recent evidence suggests that there are remarkably stable individual differences in various areas, including theory-of-mind reasoning (Wellman, Lopez-Duran, LaBounty, & Hamilton, 2008), mathematics reasoning (Halberda, Mazzocco, & Feigenson, 2008; see Opfer & Siegler, Chapter 30), and processing speed (Rose, Feldman, & Wallace, 1988). For example, individual differences at age 14 in children's ability to discriminate large, uncountable sets of dots in a visual attention task correlate highly with scores on standardized mathematics achievement tests going back to kindergarten (Halberda et al., 2008). These data suggest that differences found in infancy or early childhood relate to more complex tasks presented later in life, and they argue for continuity across strikingly different tasks.

The importance of cultural context is also a critical issue that has not been sufficiently acknowledged in the past but is beginning to receive more attention. Research with adults demonstrates that cognitive principles once thought to be universal may reflect particular cultural values (Markus & Kitayama,

1991; Nisbett, 2003). For excellent developmental work on this topic, see Greenfield, Keller, Fuligni, and Maynard (2003), Rogoff (2003), Astuti, Solomon, and Carey (2004), and Waxman, Medin, and Ross (2007).

Effects of Training and Experience

Piaget (1964) famously proposed that training cannot affect the progression of stages; the child must be developmentally ready before he or she is capable of benefitting from experience. To some extent this is certainly true: A lesson in calculus will have no effect on a 4-year-old child. Yet it is also clear that children are much more susceptible to training effects than Piaget had originally conceived. Recent intriguing studies demonstrate that certain motor experiences have broad cognitive consequences. For example, the onset of self-propelled locomotion (either crawling or use of a walker) is associated with a host of changes, including reluctance to cross a visual cliff (Campos et al., 2000). Furthermore, providing 3-month-olds with the means to pick up objects, by giving them "sticky mittens" (i.e., Velcro-backed mittens and Velcro-backed toys), leads them not only to explore the objects more intensely (Needham, Barrett, & Peterman, 2002) but also to interpret perceived events in a more sophisticated way (i.e., seeing a reach toward an object as directed toward a particular goal, rather than just a movement of a certain trajectory; Sommerville, Woodward, & Needham, 2005).

Experience can also lead to dramatic changes in perceptual and categorical processing. For example, specific experiences can lead to a process known as perceptual narrowing. At 6 months, infants can universally discriminate all the phonemes of the world's languages (e.g., p vs. b; l vs. r). However, with experience with just their native tongue(s), by 12 months infants lose the capacity to distinguish phonemes to which they are not exposed (e.g., infants exposed to just Japanese lose the capacity to distinguish l vs. r; infants exposed to just English lose the capacity to distinguish two different t sounds in Hindi; Werker & Desjardins, 1995). Perceptual narrowing is not simply an effect of speech perception, but rather seems to be found in a broad range of perceptual abilities (Scott, Pascalis, & Nelson, 2007). One intriguing example concerns face perception, where infants at 3 months show the capacity to discriminate faces of all races equally well, but by 9 months do much better discriminating faces to which they have been exposed; for example, White

babies distinguish other White faces better than Asian faces (Kelly et al., 2007).

Early Capacities

In the prior section, we reviewed some of the evidence indicating that children are much more capable than traditional Piagetian theory would suggest. In this section we consider what children's early capacities are like. We start by discussing methodological issues. Then, because this topic is too large to review in entirety, we have selected a few key themes: infants can represent objects and events; children are not limited to concrete representations; and learning concepts entails more than forming associations. Collectively, these phenomena illustrate the surprising subtlety and sophistication of early thought.

Methodology Matters

One firmly established lesson from the past 40 years of research on children's cognitive development is that methodology matters: How one operationalizes and assesses children's thinking has a powerful influence on children's performance and on the conclusions we draw about their capacities. Time and again we see that task modifications can lead to much more sophisticated performance. Children—especially young children—face difficulty when presented with tasks that require sophisticated verbal, metacognitive, planning, or information-processing skills. For example, Piaget's object permanence task requires that infants pull a cover off a hidden object to demonstrate knowledge that an object remains in existence even if it is out of sight. This requires an ability to hold and grasp the cover, as well as sufficient memory and planning skills to keep in mind the goal of retrieving the object while focusing on the task of removing the cover. The method of tracking infants' looking-time presents a much less demanding task (all babies need do is look), and it shows that infants do indeed track the existence of hidden objects, expecting them to occupy space of a certain size (and therefore evidencing surprise when a barrier moves through the space that the object should occupy; Baillargeon, 1987). More generally, numerous experimental techniques have been devised for infants that make use of the behaviors they have at their disposal (e.g., habituation, head turning, or high-amplitude sucking techniques; Cohen & Cashon, 2006).

Moreover, in some cases infants have limitations that interfere with a conceptual understanding they in fact possess. A good example of this is again

with object permanence tasks. One clever task that Piaget devised is the "A-not-B" task, where babies see a desired object hidden repeatedly at one location (A), and then see the object hidden at a new location (B). By 10 months of age, infants successfully find the object in location A, but erroneously continue to search at location A even when they saw the desired object being hidden at B. This is known as the A-not-B error. Piaget attributed this error to an incomplete representation of the object. However, it turns out that a difficulty inhibiting well-practiced responses contributes to this error (Diamond, 1991).

Even when studying older children, it is easy to underestimate their capacity. One problem is that children have difficulty verbalizing concepts that they possess. For example, children have trouble talking about traits (Livesley & Bromley, 1973), yet they show an understanding of cross-situation consistency when tested with more directed questioning (Heyman & Gelman, 1998; Liu, Gelman, & Wellman, 2007). Another issue involves information-processing demands. For example, one study found that children's pattern of picture recall seemed to differ systematically from that of adults, with children focusing exclusively on individual items and adults focusing exclusively on the category to which the items belonged (Sloutsky & Fisher, 2004a, b). However, when the length of time each item was presented to participants is controlled, by equating the length of exposure for children and adults, the developmental difference disappeared (Wilburn & Feeney, 2008).

Although simpler methods reveal earlier capacities in children, this point does not imply that there are no developmental changes in children's cognitive abilities. At times, simplified tasks reduce the need for sophisticated performance. Similarly, earlier knowledge can be more fragile or limited than later understandings. Thus, for example, although even preschool children can perform analogical reasoning when the perceptual or causal bases for comparisons are made salient (Goswami, 1992), there are nonetheless developmental changes in the depth, complexity, and sophistication of the conceptual relations children can consider (Uttal, Gentner, Liu, & Lewis, 2008).

Infants Can Represent Objects and Events

As seen by the earlier brief review of object permanence, we now know that preverbal infants can represent objects and events, even when they are

out of sight. Not only do infants represent that an object exists, they also represent details of that object, such as its height, density, solidity (e.g., object vs. substance), and capacity to contain things (Baillargeon et al., 2009). They can link visual cues with cross-modal cues (e.g., understanding that a visually bumpy surface matches a tactiley bumpy surface; Gottfried, Rose, & Bridger, 1977). They expect objects to display unity across time and space (Spelke, 2004). They represent the distinction between animate and inanimate objects and expect only animate objects to move on their own (Opfer & Gelman, 2010). They can interpret point-light displays and distinguish animate from inanimate motion on this basis (Arterberry & Bornstein, 2002). They detect patterns in complex motion, parsing action into distinct events (Baldwin, Baird, Saylor, & Clark, 2001). And they represent the number of objects in an array and can perform simple operations of addition and subtraction (Wynn, 1992).

Preverbal infants are also capable of representing information over time; that is, they form enduring memories. Thus, upon viewing a novel event, such as a person making a special lightbox light up by pushing down on it with his forehead, 9-month-olds remember the event over a delay, reproducing the event when presented with the box 1 week later (Meltzoff, 1988). Infants can remember novel events (e.g., the steps required to form a toy rattle) over periods lasting a year or more, although the quality of the memory and the amount retained over time improve with age (Bauer, 2006).

Infants still have a tremendous amount to learn about the world around them, but they have a wealth of early-emerging capacities that make sense of experience.

Young Children Are Not Limited to Concrete Representations

A long-standing and persistent view in the developmental literature is that development entails a process of moving from concrete to abstract. On this view, young children are capable of holding only *concrete* representations and are unable to form more abstract ideas (see Simons & Keil, 1995, for review). This position, however, is belied by a variety of evidence showing that even preschool children readily reason about nonobvious, nonvisible, or abstract entities. One realm in which this is powerfully demonstrated is that of reasoning about mental states (thoughts, beliefs, desires). Although there is clear improvement over time in children's theory of

mind (Wellman, in press; Wellman & Liu, 2004), even before 2 years of age, infants implicitly recognize that a person may hold a false belief (Onishi & Baillargeon, 2005), and that a failed action nonetheless has an intended goal (Brandone & Wellman, 2009; Meltzoff, 1995).

Another demonstration of children's ability to represent abstraction is in the realm of what is called "generic knowledge" (Prasada, 2000). Although children have experience strictly with individuals (this apple, that chair), they rapidly and readily generalize from experience with individuals to abstractions regarding the category as a whole (apples, chairs; Waxman & Markow, 1995). For example, by 30 months of age, children make use of subtle linguistic cues ("Blicks drink milk" vs. "These blicks drink milk") to extend a novel property to a broad class of instances (Graham, Nayer, & Gelman, 2011). Generic concepts may even be a default representation for young children, as they seem to learn generics before more formal modes of expressing generalization such as quantified noun phrases (e.g., "all blicks;" Hollander, Gelman, & Star, 2002; Leslie, 2008).

By 3 years of age, children can use verbal information to update their representations of absent objects (Ganea, Shutts, Spelke, & DeLoache, 2007) and use abstract symbols such as maps or model representations to skillfully to navigate through space (DeLoache, 1987; Uttal, Gregg, Tan, Chamberlin, & Sines, 2001). They can reason about abstract relations, such as ownership, despite the fact that an owned object is concretely no different from a nonowned object (Blake & Harris, 2009; Gelman, Manczak, & Noles, in press; Kim & Kalish, 2009; Neary, Friedman, & Burnstein, 2009; Williamson, Jaswal, & Meltzoff, 2010). For example, they place special value on their own special toys or objects and would much rather own the original attachment object than an exact duplicate (Hood & Bloom, 2008). They believe that food or drink can contain invisible particles (Au, Sidle, & Rollins, 1993; Legare, Wellman, & Gelman, 2009), and that germs can lead to illness (Kalish, 1996; Raman, 2009). They believe that unseen internal parts or power can have causal effects (Gelman, 2003; Gottfried & Gelman, 2005; Sobel, Yoachim, Gopnik, Meltzoff, & Blumenthal, 2007). They seem to posit an invisible "essence" shared by members of a category (Gelman, 2003).

The evidence indicates that children are not limited to concrete representations. Indeed, in some cases the developmental process is from abstract to concrete rather than the reverse: The child at times

starts with a more abstract representation that then gets filled in with concrete details over time (Gelman & Williams, 1998; Simons & Keil, 1995; Wellman & Gelman, 1998).

Learning Concepts Entails More Than Forming Associations

Another important theme regarding cognitive development is that when children learn concepts (and words to express those concepts), they are not simply forming associations but instead are creating meaningful interpretations of experience (Waxman & Gelman, 2009). For example, Preissler and Carey (2004) taught 18-month-old and 24-month-old toddlers a new word (“whisk”) while showing them a photograph of the referent (namely, a photo of a whisk). After repeated experience with this name-picture association, children were presented with a test in which they were asked to pick the “whisk,” given a choice between (1) another photo that was highly similar to the one used during teaching and (2) an actual, three-dimensional whisk object. If the word-learning process is merely one of association (see e.g., Sloutsky, 2003), then children should select the choice that is more perceptually similar to the taught-upon item, namely, the picture. However, if the word-learning process entails conceptual interpretation, then children should understand the relationship to involve *reference* (not merely association) and select the object (which the picture symbolizes). When this same experiment is conducted with autistic children, they respond differently, linking the word with the picture—thus arguably treating the word as associatively linked to the object rather than representing the object (Preissler & Carey, 2005).

A further important demonstration that children do not treat word-object links as purely associative is that children attend to pragmatic, communicative cues from the speaker. If such cues are unavailable, children will not link a word to a referent. Thus, if a novel word is presented in an incidental, nonintentional manner (e.g., projected over a loudspeaker in the room in which the child is sitting), children fail to link the word to the object. The word must be spoken in an intentional, face-to-face interaction (Baldwin, 1993). Furthermore, if the speaker and the child are initially focused on different objects, children check the speaker to gauge her direction of gaze and link the word to the speaker’s focus of attention, not to their own initial focus of attention (Baldwin, 1993). Put a slightly different way,

the association between word and object is blocked when the appropriate communicative cues are not in place. Again, however, autistic children perform differently, linking the word to their own focus of attention rather than that of the speaker (Baron-Cohen, Baldwin, & Crowson, 1997). Finally, if the speaker expresses some uncertainty as to the correctness of the label, children also fail to link the word to the referent (Sabbagh & Baldwin, 2001). Again, this suggests that the word-object relationship is one of reference rather than association.

Causal Reasoning and Naïve Theories

A number of researchers have advanced the view that children’s thought can be construed as sharing important similarities with scientific theories, with a focus on domain-specific ontologies, causal processes, and coherent knowledge structures that undergo qualitative reorganizations over time (Carey, 1985; Gopnik & Wellman, 1994). Although the analogy is imperfect (e.g., young children are poor at conducting systematic scientific investigations to test their theories; Kuhn et al., 1988), the “theory theory” has been a productive framework for characterizing cognitive development. In this section, we focus on the role of ontologies and causation in children’s early reasoning, as well as the implications of the theory theory for conceptual change.

Ontologies

Ontological commitments are those that distinguish different kinds of entities that participate in distinct causal laws. For example, everyday intuition tells us that mental states are distinct from physical objects, even though the mind influences the body (and vice versa), and even though mental states ultimately have a physical basis in the brain. Despite these demonstrable links between mental and physical, we act as if mental entities and physical entities are wholly different sorts of things. Indeed, children and adults alike seem to have difficulty even considering a relation between these two domains, maintaining a strict dualism (Bloom, 2004; Notaro, Gelman, & Zimmerman, 2001; Schulz, Bonawitz, & Griffiths, 2007).

From early childhood, children show evidence for ontological commitments in at least three distinct domains: physics, psychology, and biology (Wellman & Gelman, 1998). That is, they have expectations regarding three-dimensional, bounded physical objects, expectations regarding mental states and mental activities, and expectations regarding self-moving, spontaneously growing entities. Thus, even

infants expect that objects, but not shadows, occupy space (Van de Walle, Rubenstein, & Spelke, 1998), that mental states are distinct from physical objects (Wellman, *in press*), and that people and inanimate objects engage in different patterns of object motion (Opfer & Gelman, 2010).

Causality

Infants are also highly sensitive to causal relations linking objects or events. For example, 3-month-old infants readily learn the contingency between their own action and a causal effect (e.g., kicking to make a mobile move; Rovee-Collier & Barr, 2001), and within the first year of life, infants respond with more positive affect to events that they cause versus those that are uncorrelated with their actions (Gunnar-Vongnechten, 1978; Lewis, Alessandri, & Sullivan, 1990; Watson, 1972). Seven-month-old infants also distinguish a causal physical event (e.g., a ball colliding with another ball, causing it to move) from the backward version of that same event (Leslie, 1984; but see Oakes & Cohen, 1990).

An attention to causal relations can be seen in older children's commonsense theories as well. Children place special priority on causal features, treating them as more important than other sorts of features (an effect known as the "causal status hypothesis"; Ahn & Kim, 2001). For example, if children learn that a novel animal has three features, one of which (promicin in its bones) causes the other two (thick bones, big eyes), they are more likely to classify a new instance as a member of that category if it possesses the causal feature but is missing one of the other two features, than if it possesses the other two features but is missing the causal one (Ahn, Gelman, Amsterlaw, Hohenstein, & Kalish, 2000).

Within specific theories, children display a combination of openness to new causal relations (informed by statistical patterns in the input) and expectations or biases that are informed by prior knowledge and/or ontologies. So, for example, children are highly attentive to statistical cues in the input and use them to learn about how a machine works (Gopnik & Sobel, 2000), what a new word means (Xu & Tenenbaum, 2007), or how to generalize a new property from one animal to another (Rhodes, Gelman, & Brickman, 2010). At the same time, children have a "self-agency bias," whereby they are biased interpreters of statistical evidence: They are more likely to interpret their own action as having a causal effect than the actions of another person (Kushnir, Wellman, & Gelman, 2009).

As another example, children below about 10 years of age expect that mental states cannot have physical effects, thereby resisting the notion of psychogenic illness (e.g., worrying causing a stomach ache; Schulz et al., 2007). However, if children receive repeated empirical evidence in favor of such a causal account, then even 3–1/2-year-olds can overturn this assumption (Bonawitz, Fischer, & Schulz, *in press*). Children, like scientists, make use of a combination of preexisting theoretical assumptions and empirical evidence to derive new conclusions (see Cheng & Buehner, Chapter 12).

Theory Change

One of the more far-reaching implications of treating children's concepts as organized into naive theories is what this approach suggests about conceptual change. Carey (2009, p. 18) argued that theory change in childhood is analogous to theory change in science: "human beings, alone among animals, have the capacity to create representational systems that transcend sensory representations and core cognition... [they] create new representational resources that are qualitatively different from the representations they are built from." For example, although infants have an innate capacity to represent quantities (a parallel individuation system that represents individuals and so permits solving simple addition and subtraction problems; Wynn, 1992), they do not accurately represent positive integers until 3 or 4 years of age (Carey, 2009). Other domains in which conceptual change has been argued for and studied include object kinds, social kinds, theory of mind, matter/substance, and heat/temperature.

Learning From Others

To this point, we have largely focused on the impressive capacities and knowledge of infants and children. However, cognitive development happens not just in the head of the developing individual but also through a process of social interactions with others from whom children can learn (without having to discover everything themselves; see Gelman, 2009; Rogoff, 2003; Vygotsky, 1934/1962).

Learning from others is a powerful tool for gaining knowledge about a range of topics, and especially those that are difficult to observe firsthand (Harris & Koenig, 2006). For example, during the preschool and elementary school years, children learn about how our brains are related to thinking, remembering, and personal identity (Johnson

& Wellman, 1982); they learn that the earth that we live on is a sphere despite the flat appearance of the surrounding ground (Vosniadou & Brewer, 1992); and they learn that when a living thing dies, it is often related to the breakdown or cessation of hidden internal body parts (such as a heart that stops beating; Slaughter & Lyons, 2003). Coming to these conclusions would be difficult without the provision of information (or “testimony”) from more expert adults (Harris & Koenig, 2006).

Pedagogical Stance

From infancy onward, children seem to expect that the people around them have the goal of teaching them new information (i.e., a “pedagogical” stance; Csibra & Gergely, 2009). Infants and young children interpret and learn differently from situations that they construe as pedagogical (e.g., where the adult deliberately seeks and maintains eye contact with the child) than from situations that they construe as nonpedagogical (e.g., where the adult makes no attempt to engage the child’s attention). They seem biased to interpret intentional communication as conveying information that will generalize to new contexts. For example, to use Csibra and Gergely’s example, if I demonstrate how to open a certain type of container (e.g., a milk carton), the child will assume that I am teaching him how to open containers of that kind in general. He does not need repeated demonstrations in order to reach this conclusion. Furthermore, some of children’s seeming “errors” can be reinterpreted in light of this bias. For example, consider infants’ classic A-not-B error (searching for an object in a location where it has most typically been hidden, rather than in the location where it was last hidden; Piaget, 1954). Csibra and Gergely suggest that this error reflects infants’ assumption that the adult is teaching them where the toy is supposed to be stored (i.e., they assume that the adult is teaching them something general about this toy). And indeed, consistent with this interpretation, if the experimenter does not provide communicative cues (e.g., by not first making eye contact with the infant), then the A-not-B error is greatly reduced (Topál, Gergely, Miklósi, Erdőhegyi, & Csibra, 2008).

Imitation

One important means of obtaining information from others is imitation, or copying the actions of others. It is well established that children can engage in imitation of others from earliest infancy

(Meltzoff & Moore, 1977; Meltzoff & Williamson, 2010). Furthermore, children’s imitation appears to have certain qualities that are distinct from that of other species. First, children imitate rather than merely emulate (Tomasello, 1999). That is, children attempt to copy the model’s (intentional) actions, and they do not simply learn something about the environment as a result of observing the results of the model’s actions. For example, if shown how to obtain a toy by holding a tool in a particular fashion, young children will carefully rotate the tool to imitate the precise action of the model, whereas apes will not attend to the manner in which the tool is used (Penn & Povinelli, Chapter 27; but see Whiten, McGuigan, Marshall-Pescini, & Hopper, 2009, for some evidence of imitation in chimpanzees).

Second, children overimitate rather than just focus on the relevant task dimensions (Lyons, Young, & Keil, 2007). For example, if shown how to operate a novel machine using a sequence of actions that includes both causally efficacious (e.g., removing a bolt) and causally irrelevant steps (e.g., tapping on an empty compartment), preschool children will imitate both the efficacious and irrelevant steps. Lyons et al. suggest that children do so because they presume that all the steps provided by the model are causally relevant, even in the face of seemingly contradictory evidence.

And third, children imitate intentional actions and intended actions, rather than just surface behaviors (Meltzoff, 1995). Thus, if 18-month-old children view an adult engaging in an unsuccessful attempt to extract a toy from a location, but where the intended goal is clear, they will imitate in such a way as to carry out the intended action, and not simply to carry out the observed (uncompleted) motor movements. Thus, children’s imitation does not simply involve mimicking surface features of an action, but it entails consideration of the model’s intentions. If an action is highlighted as clearly accidental rather than intentional, toddlers will selectively imitate the intentional aspect (Carpenter, Akhtar, & Tomasello, 1998). Similarly, if an unusual action (e.g., turning on a lightbox with one’s head rather than one’s hands) has no obvious causal basis, children imitate precisely what the model has done (turning it on with their head), but if the unusual action has an obvious causal basis (e.g., the model’s hands are otherwise occupied), then children imitate the more efficient means of carrying out the result (e.g., turning it on with their hands) (Gergely, Bekkering, & Király, 2002).

Evaluating Testimony

Although children learn in contexts that are intended to be instructional, they also extract a great deal of knowledge from the everyday activities taking place around them (Callanan, 2006). For example, everyday conversations provide a rich source of information about gender (Gelman, Taylor, & Nguyen, 2004) and natural kind categories (Gelman, Coley, Rosengren, Hartman, & Pappas, 1998)—even when adults do not intend to teach children specifically about these topics. In addition, children play an active role in obtaining information through conversation. By 2.5–3 years of age, children begin to actively seek knowledge from others by persistently asking causal questions of adults (Callanan & Oakes, 1992; Chouinard, 2007; Hood & Bloom, 1979). Even from this young age, children appear to be asking these “why” and “how” questions with the goal of obtaining explanations—they react with satisfaction when they get an explanation and are likely to repeat their question when they do not (Frazier, Gelman, & Wellman, 2009; see Lombrozo, Chapter 14).

However, the information provided by those around us is not always accurate, as informants may intentionally or unintentionally provide incorrect information. Again, children are not passive in this process. Children actively discriminate between different sources of information and choose from whom they should learn (or, alternatively, ignore).

Early work on children’s attention to adult “testimony” was focused on word-learning tasks. For example, Koenig, Clément, and Harris (2004) asked children to observe two individuals: one who gave correct labels for familiar objects, and one who gave incorrect labels (e.g., stating, “That’s a shoe” in reference to a ball). The children then saw the same two individuals naming a novel object with two different novel names (“That’s a mido/toma”). When asked what the object is called (“Is this a mido or a toma?”), both 3- and 4-year-olds in this study picked the label from the accurate informant over the one provided by the previously inaccurate informant.

This methodology (choosing between two informants with conflicting claims) has been extended from word learning to a variety of tasks, including (but not limited to) causal inferences (Kushnir, Wellman, & Gelman, 2009), object functions (Birch, Vauthier, & Bloom, 2008), and emulation of actions (Birch, Akmal, & Frampton, 2010). This research has also explored many factors that affect

which person’s testimony children will choose to use to guide their learning. Children robustly attend to testimony from accurate informants across a variety of situations and types of knowledge. Preschoolers track the relative accuracy of informants (Pasquini, Corriveau, Koenig, & Harris, 2007), favoring those who provide the highest proportion of accurate answers. Three- and 4-year-olds are sensitive to not only what an informant knows but also whether that informant has been allowed to use that knowledge (e.g., when the knowledgeable informant has been blindfolded; Kushnir et al., 2009). Five-year-olds expect that an individual who provides the correct labels for familiar objects will also know more about words, will know more general facts, and, interestingly, will also behave more prosocially (Brosseau-Liard & Birch, 2010). Children are even willing to override their own observations when they conflict with the testimony being offered by a previously accurate informant (Jaswal, 2010; Ma & Ganea, 2010).

Children’s ability to evaluate informants also demonstrates appropriate flexibility; information regarding the social characteristics of the informant interacts with information regarding their accuracy in the decisions children make about whom to believe. For example, children evaluate an expert as more knowledgeable than a nonexpert (Lutz & Keil, 2002). And when given the choice between a child and an adult informant, if both are accurate, 3- and 4-year-old children prefer the adult (Jaswal & Neely, 2006). However, if the adult provides inaccurate labels, preschool-aged children trust the testimony of a child over that of the inaccurate adult (Jaswal & Neely, 2006). In addition, providing trait labels (referring to an informant as “very good” or “not very good” at answering a question as opposed to “right” or “wrong”) can lead 4-year-olds to prefer an accurate informant after only one trial (Fitneva & Dunfield, 2010). Even 2-year-old children use non-verbal cues from a single instance to select confident informants over informants who appear uncertain (Birch et al., 2010). However, an open question in this research concerns developmental changes in children’s ability to critically evaluate the statements of deceptive individuals with self-serving motives (Heyman, 2008).

A larger open question for this body of research concerns how children’s trust in testimony relates to their important social relationships. We know that preschoolers ask questions and that they are capable of focusing on the accuracy of testimony (even in the

presence of other conflicting, tempting factors), but we do not yet know much about whose testimony children seek out and pay attention to within their everyday lives. A study by Corriveau et al. (2009) begins to explore this question in an interesting way. These researchers looked at the relationship between the attachment status of preschoolers and their willingness to accept their mother's versus a stranger's claims. The researchers found that children with secure attachment status preferred their mother's claims over that of the stranger (unless there were conflicting perceptual cues favoring the stranger's claim). However, regardless of the perceptual cues, children classified as having an avoidant attachment status were more likely to favor the stranger's claim over their mother's, and the children classified as reactive attachment status showed the opposite pattern (relying more on their mother's claims even with conflicting perceptual information that favored the stranger's claim).

Conclusions and Future Directions

Children's thinking is of broad interest to cognitive scientists, because it provides insight into fundamental issues concerning the building blocks of cognition, the role of experience, and how conceptual change comes about. As reviewed in this chapter, prior research provides a wealth of evidence regarding these issues. Nonetheless, many open questions remain. One set of questions concerns the early capacities in infancy that have been uncovered over the past 30 years. Increasingly sophisticated methods have revealed increasingly sophisticated understandings in preverbal infants, yet we still do not fully understand the basis of these capacities. What is innate, and what is learned rapidly during the first few months of life? Is early infancy a kind of critical period? In a related vein, what kinds of input should children receive at a young age? For example, to what extent do the first few months or years of life represent a special period during which rich exposure to different languages, faces, and so on is crucial, and to what extent can these capabilities be acquired later in development? Another puzzle raised by the findings of early capacities in infancy is why infants sometimes succeed on tasks that older children fail (e.g., theory of mind understanding; grasp of physical laws; e.g., Hood, Carey, & Prasada, 2000; Onishi & Baillargeon, 2005).

Another set of open issues stems from the finding that much of children's knowledge comes about from social interactions with others (Gelman, 2009).

Clearly, social understanding plays a major role in cognitive understanding. This then raises the question of how children determine which sources of knowledge to attend to and learn from. Children must somehow sort out the different sources of knowledge, to figure out whom to believe and trust, and whom not to believe or trust. To what extent do (and should) children learn from media sources (e.g., TV, books, videos)? Other useful endeavors would be to integrate the research on testimony with research examining the questions children ask in informal learning contexts, and to investigate how children and adults work together to build an understanding of a phenomenon (e.g., Crowley et al., 2001; Siegel, Esterly, Callanan, Wright, & Navarro, 2007; see Callanan & Valle, 2008 for a related integration).

Finally, future research would benefit from examining cognitive development from new perspectives—both new from a comparative approach and new from a methodological approach. Examining children's thought in a variety of social and cultural contexts promises to reveal new insights. New advances are being made in examining what (if anything) is unique to humans versus other species. And of course it will be crucial to understand the neurological bases of cognitive performance and cognitive change.

Acknowledgments

Support for this chapter was supported by NICHD grant HD-36043 to Susan A. Gelman. We thank Keith Holyoak and Bob Siegler for very helpful comments on a prior draft.

References

- Ahn, W., Gelman, S., Amsterlaw, J., Hohenstein, J., & Kalish, C. (2000). Causal status effect in children's categorization. *Cognition*, 76, B35–B43.
- Ahn, W., & Kim, N. S. (2001). The causal status effect in categorization: An overview. In D. L. Medin (Ed.), *The psychology of learning and motivation*, Vol. 40 (pp. 23–65). San Diego, CA: Academic Press.
- Arterberry, M., & Bornstein, M. (2002). Infant perceptual and conceptual categorization: The roles of static and dynamic stimulus attributes. *Cognition*, 86, 1–24.
- Astuti, R., Solomon, G. E. A., & Carey, S. (2004). Constraints on conceptual development: A case study of the acquisition of folkbiological and folksociological knowledge in Madagascar. *Monographs of the Society for Research in Child Development*, 69, 1–135.
- Au, T. K., Sidle, A. L., & Rollins, K. B. (1993). Developing an intuitive understanding of conservation and contamination: Invisible particles as a plausible mechanism. *Developmental Psychology*, 29, 286–299.
- Baillargeon, R. (1987). Object permanence in 3½- and 4½-month-old infants. *Developmental Psychology*, 23, 655–664.
- Bauer, P. (2006). Constructing a past in infancy: A neurodevelopmental account. *Trends in Cognitive Sciences*, 10, 175–181.

- Baillargeon, R., Li, J., Ng, W., & Yuan, S. (2009). An account of infants' physical reasoning. In A. Woodward & A. Needham (Eds.), *Learning and the infant mind* (pp. 66–116). New York: Oxford University Press.
- Baldwin, D. A. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language*, 20, 395–418.
- Baldwin, D. A., Baird, J. A., Saylor, M. M., & Clark, M. A. (2001). Infants parse dynamic action. *Child Development*, 72, 708–717.
- Baron-Cohen, S., Baldwin, D., & Crowson, M. (1997). Do children with autism use the speaker's direction of gaze strategy to crack the code of language? *Child Development*, 68, 48–57.
- Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind*. New York: Oxford University Press.
- Birch, S., Akmal, N., & Frampton, K. (2010). Two-year-olds are vigilant of others non-verbal cues to credibility. *Developmental Science*, 13, 363–369.
- Birch, S. A. J., Vauthier, S. A., & Bloom, P. (2008). Three- and four-year-olds spontaneously use others' past performance to guide their learning. *Cognition*, 107, 1018–1034.
- Blake, P., & Harris, P. (2009). Children's understanding of ownership transfers. *Cognitive Development*, 24, 133–145.
- Bloom, P. (2004). *Descartes' baby: How the science of child development explains what makes us human*. New York: Basic Books.
- Bonawitz, E.B., Fischer, A., & Schulz, L.E. (in press). Teaching the Bayesian child: Three-and-a-half-year-olds' reasoning about ambiguous evidence. *Journal of Cognition and Development*.
- Brandone, A. C., & Wellman, H. M. (2009). You can't always get what you want: Infants understand failed goal-directed actions. *Psychological Science*, 20, 85–91.
- Brosseau-Liard, P., & Birch, S. (2010). 'I bet you know more and are nicer too!': What children infer from others' accuracy. *Developmental Science*, 13, 772–778.
- Callanan, M., & Oakes, L. (1992). Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. *Cognitive Development*, 7, 213–233.
- Callanan, M. (2006). Cognitive development, culture, and conversation: Comments on Harris and Koenig's "Trust in testimony: How children learn about science and religion." *Child Development*, 77, 525–530.
- Callanan, M., & Valle, A. (2008). Co-constructing conceptual domains through family conversations and activities. In B. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 49, pp. 147–165). New York: Elsevier.
- Campos, J. J., Anderson, D. I., Barbu-Roth, M. A., Hubbard, E. M., Hertenstein, M. J., & Witherington, D. (2000). Travel broadens the mind. *Infancy*, 1, 149–219.
- Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: Bradford Books, MIT Press.
- Carey, S. (2009). *The origin of concepts*. New York: Oxford University Press.
- Carpenter, M., Akhtar, N., & Tomasello, M. (1998). Fourteen-through 18-month-old infants differentially imitate intentional and accidental actions. *Infant Behavior and Development*, 21, 315–330.
- Case, R. (1992). The role of the frontal lobes in the regulation of cognitive development. *Brain and Cognition*, 20, 51–73.
- Case, R. (1995). Capacity-based explanations of working memory growth: A brief history and reevaluation. In F. E. Weinert and W. Schneider (Eds.), *Memory performance and competencies: Issues in growth and development* (pp. 23–44). Hillsdale, NJ: Erlbaum.
- Chi, M. (1978). Knowledge structures and memory development. In R. Siegler (Ed.), *Children's thinking: What develops?* (pp. 73–96). Hillsdale, NJ: Erlbaum.
- Chouinard, M. M. (2007). Children's questions: A mechanism for cognitive development. *Monographs of the Society for Research in Child Development*, 72(1, Serial No. 286).
- Cohen, L., & Cashon, C. (2006). *Infant cognition. Handbook of child psychology. Vol 2: Cognition, perception, and language* (6th ed., pp. 214–251). Hoboken, NJ: Wiley.
- Corriveau, K., Harris, P., Meins, E., Fernyhough, C., Arnott, B., Elliott, L.,...de Rosnay, M. (2009). Young children's trust in their mother's claims: Longitudinal links with attachment security in infancy. *Child Development*, 80, 750–761.
- Crowley, K., Callanan, M.A., Jipson, J.L., Galco, J., Topping, K., & Shrager, J. (2001). Shared scientific thinking in everyday parent-child activity. *Science Education*, 85, 712–732.
- Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, 13, 148–153.
- DeLoache, J. (1987). Rapid change in the symbolic functioning of very young children. *Science*, 238, 1556–1557.
- Diamond, A. (1991). Neuropsychological insights into the meaning of object concept development. In S. Carey & R. Gelman (Eds.), *The epigenesis of mind: Essays on biology and cognition* (pp. 67–110). Hillsdale, NJ: Erlbaum.
- Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, 8, 307–314.
- Fishbein, H., Lewis, S., & Keiffer, K. (1972). Children's understanding of spatial relations: Coordination of perspectives. *Developmental Psychology*, 7, 21–33.
- Fitneva, S., & Dunfield, K. (2010). Selective information seeking after a single encounter. *Developmental Psychology*, 46, 1380–1384.
- Flavell, J. H. (1963). *The developmental psychology of Jean Piaget*. New York: van Nostrand.
- Flavell, J., Everett, B., Croft, K., & Flavell, E. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1–Level 2 distinction. *Developmental Psychology*, 17, 99–103.
- Frazier, B. N., Gelman, S. A., & Wellman, H. M. (2009). Preschoolers' search for explanatory information within adult-child conversation. *Child Development*, 80, 1592–1611.
- Ganea, P., Shutts, K., Spelke, E., & DeLoache, J. (2007). Thinking of things unseen: Infants' use of language to update mental representations. *Psychological Science*, 18, 734–739.
- Gelman, R., & Williams, E. (1998). Enabling constraints for cognitive development and learning: Domain specificity and epigenesis. In D. Kuhn & R. Siegler (Eds.), *Cognition, perception and language. Vol. 2: Handbook of Child Psychology* (5th ed., pp. 575–630). New York: Wiley.
- Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. New York: Oxford University Press.
- Gelman, S. A. (2009). Learning from others: Children's construction of concepts. *Annual Review of Psychology*, 60, 115–140.
- Gelman, S. A., Coley, J. D., Rosengren, K., Hartman, E., & Pappas, A. (1998). Beyond labeling: The role of maternal input in the acquisition of richly-structured categories. *Monographs of the Society for Research in Child Development*, 63(1, Serial No. 253).
- Gelman, S. A., Manczak, E. M., & Noles, N. S. (in press). The non-obvious basis of ownership: Preschool children trace the history and value of owned objects. *Child Development*.

- Gelman, S. A., & Nokes, N. S. (2011). Domains and naïve theories. *Wiley Interdisciplinary Reviews Cognitive Science*, 2(5), 490–502.
- Gelman, S. A., Taylor, M. G., & Nguyen, S. (2004). Mother-child conversations about gender: Understanding the acquisition of essentialist beliefs. *Monographs of the Society for Research in Child Development*, 69(1, Serial No. 275).
- Gergely, G., Bekkering, H., & Király, I. (2002). Rational imitation in preverbal infants. *Nature*, 415, 68–73.
- Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world*. New York: Oxford University Press.
- Ginsburg, H., & Opper, S. (1969). *Piaget's theory of intellectual development: An introduction*. Englewood Cliffs, NJ: Prentice-Hall.
- Gopnik, A., & Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends in Cognitive Sciences*, 8, 371–377.
- Gopnik, A., & Sobel, D. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development*, 71, 1205–1222.
- Gopnik, A., & Wellman, H. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 257–293). New York: Cambridge University Press.
- Goswami, U. (1992). *Analogical reasoning in children*. Hillsdale, NJ: Erlbaum.
- Gottfried, A. W., Rose, S. A., & Bridger, W. H. (1977). Cross-modal transfer in human infants. *Child Development*, 48, 118–123.
- Gottfried, G. M., & Gelman, S. A. (2005). Developing domain-specific causal explanatory frameworks: The role of insides and immanence. *Cognitive Development*, 20, 137–158.
- Gould, S. J. (1983). *Hen's teeth and horse's toes*. New York: W. W. Norton.
- Graham, S. A., Nayer, S. L., & Gelman, S. A. (2011). Two-year-olds use the generic/non-generic distinction to guide their inferences about novel kinds. *Child Development*, 82, 492–507.
- Greenfield, P. M., Keller, H., Fuligni, A., & Maynard, A. (2003). Cultural pathways through universal development. *Annual Review of Psychology*, 54, 461–490.
- Gunnar-vonGnechten, M. (1978). Changing a frightening toy into a pleasant toy by allowing the infant to control its actions. *Developmental Psychology*, 14, 157–162.
- Halberda, J., Mazzocco, M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, 455, 665–669.
- Halford, G., Cowan, N., & Andrews, G. (2007). Separating cognitive capacity from knowledge: A new hypothesis. *Trends in Cognitive Sciences*, 11, 236–242.
- Harris, P., & Koenig, M. (2006). Trust in testimony: How children learn about science and religion. *Child Development*, 77, 505–524.
- Heyman, G. (2008). Children's critical thinking when learning from others. *Current Directions in Psychological Science*, 17, 344–347.
- Heyman, G., & Gelman, S. (1998). Young children use motive information to make trait inferences. *Developmental Psychology*, 34, 310–321.
- Hollander, M. A., Gelman, S. A., & Star, J. (2002). Children's interpretation of generic noun phrases. *Developmental Psychology*, 38, 883–894.
- Hood, B., & Bloom, P. (2008). Children prefer certain individuals over perfect duplicates. *Cognition*, 106, 455–462.
- Hood, B., Carey, S., & Prasada, S. (2000). Predicting the outcomes of physical events: Two-year-olds fail to reveal knowledge of solidity and support. *Child Development*, 71, 1540–1554.
- Hood, L., & Bloom, L. (1979). What, when, and how about why: A longitudinal study of early expressions of causality. *Monographs of the Society for Research in Child Development*, 44(6, Serial No. 181).
- Jaswal, V. (2010). Believing what you're told: Young children's trust in unexpected testimony about the physical world. *Cognitive Psychology*, 61, 248–272.
- Jaswal, V., & Neely, L. (2006). Adults don't always know best: Preschoolers use past reliability over age when learning new words. *Psychological Science*, 17, 757–758.
- Johnson, C., & Wellman, H. (1982). Children's developing conceptions of the mind and brain. *Child Development*, 53, 222–234.
- Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, 21, 60–99.
- Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.
- Kalish, C. (1996). Preschoolers' understanding of germs as invisible mechanisms. *Cognitive Development*, 11, 83–106.
- Kanwisher, N.G., & Yovel G. (2009). Cortical specialization for face perception in humans. In J. T. Cacioppo & G. G. Bernston (Eds.), *Handbook of neuroscience for the behavioral sciences* (pp. 3–20). New York: Wiley.
- Keil, F. (1981). Constraints on knowledge and cognitive development. *Psychological Review*, 88, 197–227.
- Keil, F., Smith, W., Simons, D., & Levin, D. (1998). Two dogmas of conceptual empiricism: Implications for hybrid models of the structure of knowledge. *Cognition*, 65, 103–135.
- Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Ge, L., & Pascalis, O. (2007). The other-race effect develops during infancy: Evidence of perceptual narrowing. *Psychological Science*, 18, 1084–1089.
- Keysar, B. (2007). Communication and miscommunication: The role of egocentric processes. *Intercultural Pragmatics*, 4, 71–84.
- Keysar, B., Barr, D., Balin, J., & Brauner, J. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, 11, 32–38.
- Kim, S., & Kalish, C. W. (2009). Children's ascriptions of property rights with changes in ownership. *Cognitive Development*, 24, 322–336.
- Koenig, M., Clément, F., & Harris, P. (2004). Trust in testimony: Children's use of true and false statements. *Psychological Science*, 15, 694–698.
- Kuhl, P. K., Tsao, F.-M., & Liu, H.-M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences USA*, 100, 9096–9101.
- Kuhn, D., Amsel, E., O'Loughlin, M., Schauble, L., Leadbeater, B., & Yotive, W. (1988). *The development of scientific thinking skills*. San Diego, CA: Academic Press.
- Kushnir, T., Wellman, H. M., & Gelman, S. A. (2009). A self-agency bias in preschoolers' causal inferences. *Developmental Psychology*, 45, 597–603.
- Legare, C. H., Wellman, H. M., & Gelman, S. A. (2009). Evidence for an explanation advantage in naive biological reasoning. *Cognitive Psychology*, 58, 177–194.

- Leslie, A. (1984). Spatiotemporal continuity and the perception of causality in infants. *Perception*, 13, 287–305.
- Leslie, S. J. (2008). Generics: Cognition and acquisition. *The Philosophical Review*, 117, 1–49.
- Lewis, M., Alessandri, S. M., & Sullivan, M. W. (1990). Violation of expectancy, loss of control, and anger expressions in young infants. *Developmental Psychology*, 26, 745–751.
- Liu, D., Gelman, S. A., & Wellman, H. M. (2007). Components of young children's trait understanding: Behavior-to-trait inferences and trait-to-behavior predictions. *Child Development*, 78, 1543–1558.
- Livesley, W., & Bromley, D. (1973). *Person perception in childhood and adolescence*. Oxford, England: Wiley.
- Lutz, D., & Keil, F. (2002). Early understanding of the division of cognitive labor. *Child Development*, 73, 1073–1084.
- Lyons, D. E., Young, A. G., & Keil, F. C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences USA*, 104, 19751–19756.
- Ma, L., & Ganea, P. (2010). Dealing with conflicting information: Young children's reliance on what they see versus what they are told. *Developmental Science*, 13, 151–160.
- Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, 98, 224–253.
- Meltzoff, A. (1988). Infant imitation after a 1-week delay: Long-term memory for novel acts and multiple stimuli. *Developmental Psychology*, 24, 470–476.
- Meltzoff, A. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, 31, 838–850.
- Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, 198, 75–78.
- Meltzoff, A. N., & Williamson, R. A. (2010). The importance of imitation for theories of social-cognitive development. In G. Bremner & T. Wachs (Eds.), *Handbook of infant development* (2nd ed., pp. 345–364). Oxford: Wiley-Blackwell.
- Neary, K., Friedman, O., & Burnstein, C. (2009). Preschoolers infer ownership from "control of permission." *Developmental Psychology*, 45, 873–876.
- Needham, A., Barrett, T., & Peterman, K. (2002). A pick me up for infants' exploratory skills: Early simulated experiences reaching for objects using 'sticky' mittens enhances young infants' object exploration skills. *Infant Behavior and Development*, 25, 279–295.
- Nisbett, R. E. (2003). *The geography of thought: How Asians and Westerners think differently...and why*. New York: The Free Press.
- Notaro, P. C., Gelman, S. A., & Zimmerman, M. A. (2001). Children's understanding of psychogenic bodily reactions. *Child Development*, 72, 444–459.
- Oakes, L. M., & Cohen L. B. (1990). Infant perception of a causal event. *Cognitive Development*, 5, 193–207.
- O'Neill, D. (1996). Two-year-old children's sensitivity to a parent's knowledge state when making requests. *Child Development*, 67, 659–677.
- Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308, 255–258.
- Opfer, J. E., & Gelman, S. A. (2010). Development of the animate-inanimate distinction. In U. Goswami (Ed.), *Wiley-Blackwell handbook of childhood cognitive development* (pp. 213–238). Cambridge, UK: Blackwell.
- Oztop, E., Bradley N. S., & Arbib, M. A. (2004). Infant grasp learning: A computational model. *Experimental Brain Research*, 158, 480–503.
- Pasquini, E., Corriveau, K., Koenig, M., & Harris, P. (2007). Preschoolers monitor the relative accuracy of informants. *Developmental Psychology*, 43, 1216–1226.
- Piaget, J. (1954). *The construction of reality in the child*. New York: Basic Books.
- Piaget, J. (1964). Development and learning. In R. E. Ripple & V. N. Rockcastle (Eds.), *Piaget rediscovered* (pp. 7–20). Ithaca, NY: Cornell University.
- Piaget, J. (2000). Piaget's theory. In K. Lee (Eds.), *Childhood cognitive development: The essential readings* (pp. 33–47). Malden, MA: Blackwell Publishing.
- Piaget, J., & Inhelder, B. (1956). *The child's conception of space*. London: Routledge.
- Pinker, S. (1994). *The language instinct*. New York: W. Morrow.
- Prasada, S. (2000). Acquiring generic knowledge. *Trends in Cognitive Sciences*, 4, 66–72.
- Preissler, M. A., & Carey, S. (2004). Do both pictures and words function as symbols for 18- and 24-month-old children? *Journal of Cognition and Development*, 5, 185–212.
- Preissler, M. A., & Carey, S. (2005). What is the role of intentional inference in word learning? Evidence from autism. *Cognition*, 97, B13–B23.
- Rakison, D., & Lupyan, G. (2008). Developing object concepts in infancy: An associative learning perspective. *Monographs of the Society for Research in Child Development*, 73, 1–110.
- Rakoczy, H., Warneken, F., & Tomasello, M. (2007). "This way!" "No! That way!"—3-year olds know that two people can have mutually incompatible desires. *Cognitive Development*, 22, 47–68.
- Raman, L. (2009). Can we get sick if we want to? Children's and adults' recognition of intentionality in the origins of illness and injuries. *British Journal of Psychology*, 100, 729–751.
- Repacholi, B., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology*, 33, 12–21.
- Rhodes, M., Gelman, S. A., & Brickman, D. (2010). Children's attention to sample composition in learning, teaching, and discovery. *Developmental Science*, 13, 421–429.
- Rogoff, B. (2003). *The cultural nature of human development*. New York: Oxford University Press.
- Rose, S. A., Feldman, J. F., & Wallace, I. F. (1988). Individual differences in infants' information processing: Reliability, stability, and prediction. *Child Development*, 59, 1177–1197.
- Rovee-Collier, C., & Barr, R. (2001). Infant learning and memory. In G. Bremner & A. Fogel (Eds.), *Blackwell handbook of infant development* (pp. 139–168). Malden, MA: Blackwell.
- Sabbagh, M. A., & Baldwin, D. A. (2001). Learning words from knowledgeable versus ignorant speakers: Links between preschoolers' theory of mind and semantic development. *Child Development*, 72, 1054–1070.
- Schulz, L. E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can being scared cause tummy aches? Naïve theories, ambiguous evidence, and preschoolers' causal inferences. *Developmental Psychology*, 43, 1124–1139.
- Scott, L., Pascalis, O., & Nelson, C. (2007). A domain-general theory of the development of perceptual discrimination. *Current Directions in Psychological Science*, 16, 197–201.
- Siegel, D., Esterly, J., Callanan, M., Wright, R., & Navarro, R. (2007). Conversations about science across activities in

- Mexican-descent families. *International Journal of Science Education*, 29, 1447–1466.
- Sieger, R., & Crowley, K. (1991). The microgenetic method: A direct means for studying cognitive development. *American Psychologist*, 46, 606–620.
- Simons, D., & Keil, F. (1995). An abstract to concrete shift in the development of biological thought: The insides story. *Cognition*, 56, 129–163.
- Slaughter, V., & Lyons, M. (2003). Learning about life and death in early childhood. *Cognitive Psychology*, 46, 1–30.
- Sloutsky, V. M. (2003). The role of similarity in the development of categorization. *Trends in Cognitive Sciences*, 7, 246–251.
- Sloutsky, V. M., & Fisher, A. V. (2004a). When development and learning decrease memory: Evidence against category-based induction in children. *Psychological Science*, 15, 553–558.
- Sloutsky, V. M., & Fisher, A. V. (2004b). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General*, 133, 166–188.
- Sobel, D. M., Yoachim, C. M., Gopnik, A., Meltzoff, A. N., & Blumenthal, E. J. (2007). The blicket within: Preschoolers' inferences about insides and causes. *Journal of Cognition and Development*, 8, 159–182.
- Sommerville, J. A., Woodward, A. L., & Needham, A. (2005). Action experience alters 3-month-old infants' perception of others' actions. *Cognition*, 96, B1–B11.
- Spelke, E. S. (2004). Core knowledge. In N. Kanwisher & J. Duncan (Eds.), *Attention and Performance: Functional neuroimaging of visual cognition* (Vol. 20, pp. 29–56). Oxford, England: Oxford University Press.
- Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.
- Topál, J., Gergely, G., Miklósi, Á., Erdőhegyi, Á., & Csiba, G. (2008). Infant perseverative errors are induced by pragmatic misinterpretation. *Science*, 321, 1831–1834.
- Treiman, R., & Kessler, B. (2007). Learning to read. In M. G. Gaskell (Ed.), *Oxford handbook of psycholinguistics* (pp. 657–666). Oxford, England: Oxford University Press.
- Uttal, D., Gentner, D., Liu, L., & Lewis, A. (2008). Developmental changes in children's understanding of the similarity between photographs and their referents. *Developmental Science*, 11, 156–170.
- Uttal, D. H., Gregg, V. H., Tan, L. S., Chamberlin, M. H., & Sines, A. (2001). Connecting the dots: Children's use of a systematic figure to facilitate mapping and search. *Developmental Psychology*, 37, 338–350.
- Van Boven, L., & Loewenstein, G. (2003). Social projection of transient drive states. *Personality and Social Psychology Bulletin*, 29, 1159–1168.
- Van de Walle, G., Rubenstein, J., & Spelke, E. (1998). Infant sensitivity to shadow motions. *Cognitive Development*, 13, 387–419.
- von Hofsten, C. (2007). Action in development. *Developmental Science*, 10, 54–60.
- Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: a study of conceptual change in childhood. *Cognitive Psychology*, 24, 535–585.
- Vygotsky, L. S. (1934/1962). *Thought and language*. Cambridge, MA: MIT Press.
- Watson, J. (1972). Smiling, cooing, and 'the game.' *Merrill-Palmer Quarterly*, 18, 323–339.
- Waxman, S. R., & Gelman, S. A. (2009). Early word-learning entails reference, not merely associations. *Trends in Cognitive Sciences*, 13, 258–263.
- Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12-month-old infants. *Cognitive Psychology*, 29, 257–302.
- Waxman, S. R., Medin, D. L., & Ross, N. (2007). Folkbiological reasoning from a cross-cultural developmental perspective: Early essentialist notions are shaped by cultural beliefs. *Developmental Psychology*, 43, 294–308.
- Wellman, H. M. (in press). Developing a theory of mind. In U. Goswami (Ed.), *Handbook of childhood cognitive development* (2nd ed.). Oxford, England: Blackwell.
- Wellman, H. M., & Gelman, S. A. (1998). Knowledge acquisition. In D. Kuhn & R. Siegler (Eds.), *Handbook of child psychology: Cognitive development* (5th ed., pp. 523–573). New York: Wiley.
- Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, 75, 523–541.
- Wellman, H. M., Lopez-Duran, S., LaBounty, J., & Hamilton, B. (2008). Infant attention to intentional action predicts preschool theory of mind. *Developmental Psychology*, 44, 618–623.
- Werker, J. F., & Desjardins, R. N. (1995). Listening to speech in the 1st year of life: Experiential influences on phoneme perception. *Current Directions in Psychological Science*, 4, 76–81.
- Whiten, A., McGuigan, N., Marshall-Pescini, S., & Hopper, L. M. (2009). Emulation, imitation, over-imitation, and the scope of culture for child and chimpanzee. *Philosophical Transactions of the Royal Society B*, 364, 2417–2428.
- Wilburn, C., & Feeney, A. (2008). Do development and learning really decrease memory? On similarity and category-based induction in adults and children. *Cognition*, 106, 1451–1464.
- Williamson, R., Jaswal, V., & Meltzoff, A. (2010). Learning the rules: Observation and imitation of a sorting strategy by 36-month-old children. *Developmental Psychology*, 46, 57–65.
- Woodward, A., & Needham, A. (Eds.) (2009). *Learning and the infant mind*. New York: Oxford University Press.
- Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, 358, 749–750.
- Xu, F., & Tenenbaum, J. (2007). Word learning as Bayesian inference. *Psychological Review*, 114, 245–272.

The Human Enigma

Derek C. Penn and Daniel J. Povinelli

Abstract

Why is there such an enormous gap between human and nonhuman minds? Humans have been asking themselves this question for millennia. But if anything, the question has only become more enigmatic since Darwin and the genetic revolution. In the present chapter, we review the various answers that have been proposed to this question in recent years—from a “language instinct” and a “Theory of Mind” to the “massively modular” hypothesis—and argue that none of them provides a satisfactory solution to the enigma of the human mind.

Key Words: human mind, evolution, higher order relational reasoning, language, Theory of Mind, big brain, social intelligence hypothesis, language instinct

In most respects, we’re a rather unimpressive animal. We’re slow, weak, flat-footed, fragile, and flabby. We have big brains. But our social groups are mired in pointless conflict. Our eyesight and hearing pale in comparison to many bird-brained species. We can barely smell ourselves, let alone predators or prey. But it’s been a few billion years since life first evolved on this planet. And in all that time, we’re the only ones that have ever figured out how to light fires, sew, paint, call meetings, throw dinner parties, and send our children to school. Regardless of whether we want to admit it, there’s something remarkable about the human mind.

Of course, our species’ cognitive ability is hardly the only feature that distinguishes us from other animals. Compared to the rest of the great apes, we are also notable for our hairless sweaty skin, the unusual dexterity of our hands, the unique morphology of our eyes, our ability to swim, the prolonged helplessness of our children, and the peculiar biology of our sialic-acid-recognizing proteins, to mention only a few examples among many. But all these other

differences between human and nonhuman apes pale in comparison to the discontinuity between human and nonhuman cognition.

So here’s the enigma, perhaps the most profound enigma in cognitive science today: *Why is the gap between human and nonhuman minds so enormous?*

Humans have been asking themselves this question for a very long time—at least since the ancient Greeks and probably a lot longer than that. But if anything, the question has become even more enigmatic over the last century. Ever since Darwin, we have known that there is a profound biological continuity between humans and every other species. Indeed, we now know that the human species split off from other ape lineages just 5–6 million years ago—the blink of an eye on an evolutionary timescale. And we know that the vast majority of our DNA is identical to that of other living apes (Chimpanzee Sequencing and Analysis Consortium, 2005; Varki & Altheide, 2005). The more we learn about the biology of the human species, the more enigmatic our peculiar mental abilities become.

The striking gap between human and nonhuman cognition has only become more intriguing given what we have learned about nonhuman minds over the last 40-odd years (for a general overview, see Shettleworth, 2010). We now know, for example, that tool making is not a feat unique to chimpanzees—or even primates (cf. McGrew, 1992; Tomasello & Call, 1997). Corvids have turned out to be some of the most proficient tool makers in the animal kingdom (Bluff, Weir, Rutz, Wimpenny, & Kacelnik, 2007; Clayton & Emery, 2005). Species as unassuming as little African cichlid fish and great tits have turned out to behave in ways that seem, at least at first blush, to require “logical” reasoning (Groseclose, Clement, & Fernald, 2007; Peake, Terry, McGregor, & Dabelsteen, 2002). The idea that humans are the only animal with minds has been soundly discredited. Yet we are the only ones with nose rings, iPhones, and baskets. Why?

Over the last decade, scientists and philosophers have proposed any number of explanations for what makes the human mind unique, ranging from symbols (Deacon, 1997) to singing (Mithen, 2006). In this chapter, we examine the most prominent of these hypotheses and argue that none of them provides a satisfactory solution to the enigma of the human mind.

It's Not Just the Size of Our Brains

The human brain has grown enormously since our lineage separated from that of other apes (for a general overview, see Striedter, 2005). And one of the most popular hypotheses in neuroscience and anthropology over the course of the 20th century was that the difference between human and nonhuman cognition could be attributed largely to a difference in the size of our brains. This hypothesis fit well with Darwin's claim that the difference between human and non-human minds is “one of degree and not of kind” (Darwin, 1871). Indeed, Darwin was a proponent of the basest form of this hypothesis: the notion that putative differences in intelligence between human races could also be attributed to relative brain size:

The belief that there exists in man some close relation between the size of the brain and the development of the intellectual faculties is supported by the comparison of the skulls of savage and civilized races, of ancient and modern people, and by analogy of the whole vertebrate series. (Darwin, 1871 chapter II, p. 42)

The absolute size of an animal's brain, however, has turned out to be a very poor indication of an

animal's or human's intelligence (Roth & Dicke, 2005). Monkeys have brains that are much smaller, in absolute terms, than those of ungulates, although few researchers would claim that ungulates are the more intelligent species. Bird-brained species such as crows are some of the smartest animals on the planet (Seed, Emery, & Clayton, 2009). And the 1.35 kg brain of *Homo sapiens* is dwarfed by the brains of elephants and many cetaceans. Even within our own species, very little of the differences in intelligence among individuals can be attributed to differences in brain size (Rushton & Ankney, 2009; Schoenemann, Budinger, Sarich, & Wang, 2000).

Some researchers argue that it is the size of an animal's brain relative to its body, that is, its “encephalization,” that is the critical metric (see Dunbar & Shultz, 2007 for a recent review). The human brain, it is true, is four to five times larger than would be expected for an average mammal of the same size. Primates tend to have larger brains relative to body size than other mammals. And within the hominid (ape) line, the size of the human brain is an outlier: An average human brain weighs about 1,300 grams, and an average chimp brain weighs about 400 grams.

One prominent hypothesis argues that the driving force behind the evolution of larger relative brains is the demands of living within complex, often antagonistic, social relationships (Dunbar, 1993, 1998). Byrne and Whiten originally coined the term “Machiavellian intelligence” to refer to the cognitive demands of surviving among antagonistic social relationships, and they argued that the increase in relative brain size across primate lineages correlates well with a number of metrics of social complexity, including overall group size, number of females, frequency of coalitions, and frequency of incidents of social learning and tactical deception (Byrne, 1993, 1994, 1996; Byrne & Whiten, 1988). Recently, proponents of the “social brain” hypothesis, as it is now known, have argued that complex dyadic relationships, especially pair-bonds among parents, may be an even more powerful factor (Dunbar & Shultz, 2007).

But there are many reasons why encephalization has not fared much better than absolute brain size as an explanation for human cognitive uniqueness (Gazzaniga, 2008; Roth & Dicke, 2005; Schoenemann, 2006). First off, even in our own lineage, the relationship between brain size and cognitive ability is a complex story. Our ancestors had big brains for millions of years without any great leap forward in their cognitive abilities. Neanderthals probably had a

higher encephalization quotient than *Homo sapiens*, but there is little doubt that *Homo sapiens* were their intellectual superiors. Indeed, the size of *Homo sapiens'* brain has actually decreased by about 150 cc over the course of our species' history (Rightmire, 2004).

Even among other animal species, the relationship between encephalization and social complexity is spotty at best. Capuchin monkeys, for example, have higher encephalization quotients than chimpanzees and gorillas. It turns out that there is no association between sociality and encephalization across extant Carnivora (i.e., cats, dogs, bears, weasels, and their relatives). And thus there is little support for the hypothesis that sociality was the causal agent for increased encephalization in mammals in general (Finarelli & Flynn, 2007, 2009).

Many proponents of the social brain hypothesis focused on the relative size of the neocortex, arguing that this was the area that was disproportionately expanded in primates. And it is true that we have a bizarrely enlarged neocortex relative to other apes (but see Preuss, 2004; cf. Semendeferi, Armstrong, Schleicher, Zilles, & Van Hoesen, 2001). But this fact alone does not tell us much about why a big neocortex makes us so smart in our species-specific ways. Explaining human cognitive uniqueness requires a much more detailed understanding of the structural and functional differences between human and nonhuman brains.

Numerous features of the human neocortex have attracted researchers' attention in recent years, including the disproportionate amount of white matter in the human prefrontal cortex (Schoenemann, Sheehan, & Glotzer, 2005), the lateral specialization between our cerebral hemispheres (Gazzaniga, 2000), the reorganization of the human temporal lobe (Rilling & Seligman, 2002), the dynamics of neural synchrony (Uhlhaas & Singer, 2006; Varela, Lachaux, Rodriguez, & Martinerie, 2001), and the role played by our prefrontal cortices, particularly our anterior prefrontal cortices, in uniquely human forms of relational reasoning (Amati & Shallice, 2007; Cho et al., 2010; Ramnani & Owen, 2004; Semendeferi et al., 2001). But at the moment, we are still a long way from understanding how these neural-level traits conspire to produce the computations that make human cognition unique or even, for that matter, what those uniquely human computations are. And as Marr (1982) taught, there's not much point understanding the neural-level details of a cognitive system if you don't know what computations that cognitive system is doing.

It's Not Only Our Words

The most popular explanation today for human cognitive uniqueness is also the oldest one on the books: language. Ever since the ancient Greeks, philosophers have argued that the difference between human and nonhuman minds must be due to our unique capacity for language. And a large number of philosophers and scientists still claim that language is at the heart of "what makes us so smart" (Bermudez, 2003; Bickerton, 2009; Carruthers, 2002; Deacon, 1997; Dennett, 1995; Diamond, 1992; Dupre, 2003; Gentner, 2003; Mithen, 1996; Premack, 2004). Dennett (1996, p.17) has staked out the extreme version of this hypothesis: "Perhaps the kind of mind you get when you add language to it is so different from the kind of mind you can have without language that calling them both minds is a mistake."

Language clearly plays a crucial role in "extending" human cognition (Clark, 2001, 2006). The most obvious way in which human language enables human cognition is by giving us the ability to employ symbolic tokens to represent features, entities, and ideas, not just for the purposes of communicating with others but also for thinking novel thoughts by ourselves (see Carruthers, 2002). Gentner and colleagues, for example, have shown that relational "labels" play an instrumental role in facilitating children's ability to grasp relational similarities and potential analogies (Gentner & Christie, 2008; Gentner & Kurtz, 2005; Loewenstein & Gentner, 2005; and see Holyoak, Chapter 13). The ability to use symbolic tokens to count and perform mathematical calculations is another obvious example of the unique role that symbolic tokens play in human cognition (Bloom & Wynn, 1997; Dehaene, 1997; see Gleitman & Papafragou, Chapter 28). And many researchers have shown that the development of normal human social cognition is heavily influenced by a child's linguistic environment. Deaf children raised by hearing parents, for example, tend to be "late signers" and show persistent deficits in tasks that require them to reason about others' psychological states relative to deaf children raised by deaf parents (Peterson & Siegal, 1995, 2000; Siegal, Varley, & Want, 2001). Normal human cognition clearly relies on normal language development (Baldo et al., 2005).

But although language enables and catalyzes normal human cognition, there is compelling evidence that the human mind is nevertheless distinctively human even in the absence of natural language sentences (Bloom, 2000b; Garfield, Peterson, & Perry,

2001; Siegal et al., 2001; Varley, Siegal, & Want, 2001). Varley and Siegal (2000), for example, studied the cognitive abilities of an agrammatic aphasic man who was incapable of producing or comprehending sentences and whose vocabulary was essentially limited to perceptual nouns. In particular, he had lost all his vocabulary for mentalistic entities such as “beliefs” and “wants.” Yet this patient continued to take care of the family finances and passed a battery of causal reasoning and Theory of Mind (ToM) tests (see also Siegal et al., 2001; Varley, Klessinger, Romanowski, & Siegal, 2005; Varley et al., 2001). Another example: although late-signing deaf children’s cognitive abilities may not be “normal,” they nevertheless manifest grammatical, logical, and causal reasoning abilities far beyond those of any nonhuman subject (Peterson & Siegal, 2000). And the many remarkable cases of congenitally deaf children spontaneously “inventing” gestural languages with hierarchical and compositional structure provide further confirmation that there is something unique about the human mind even in the absence of normal linguistic enculturation (Goldin-Meadow, 2003; Sandler, Meir, Padden, & Aronoff, 2005; Senghas, Kita, & Ozyurek, 2004).

The cognitive accomplishments of humans deprived of normal linguistic enculturation are all the more striking in contrast to the failures of even the most well-trained animal to master anything even remotely resembling a human language. Over the last 35 years, comparative researchers have tried mightily to teach nonhuman animals of a variety of taxa to use and/or comprehend language-like symbol systems (Herman, 1986; Herman, Richards, & Wolz, 1984; Kaminski, Call, & Fischer, 2004; Pepperberg, 2002; Savage-Rumbaugh, Shanker, & Taylor, 1998; Schusterman & Gisiner, 1989; Schusterman & Krieger, 1986). Many of these animals have experienced protracted periods of enculturation that rival those of modern (coddled) human children. And the stars of these animal language projects have indeed been able to approximate certain superficial aspects of human language, including the ability to associate arbitrary sounds, tokens, and gestures with external objects, properties, and actions and a rudimentary sensitivity to the order in which these “symbols” appear.

But the most striking thing about the achievements of all these animals is how rudimentary they are relative to those of human children. Even after decades of exhaustive training, no nonhuman animal has demonstrated a mastery of abstract grammatical

categories, closed-class items, hierarchical syntactic structures, or any of the other defining features of a human language (cf. Kako, 1999). There is no evidence that even the most exceptional animal understands “words” with anything like the flexibility or systematicity of the average human child (Bloom, 2000a, 2004). If any human child learned words the way animals do, the child’s parents would run screaming to their local pediatrician.

If the history of animal language research demonstrates nothing else, it demonstrates that you cannot create a human mind simply by taking a nonhuman one and teaching it to use language-like symbols. There must be substantive differences between human and nonhuman minds that allow the former, but not any of the latter, to master grammatically structured languages to begin with. Darwin (1871, p. 57) put it perfectly: “the mental powers of some early progenitor of man must have been more highly developed than in any existing ape, before even the most imperfect form of speech could have come into use.”

It’s Not Our “Language Instinct”

Ever since Noam Chomsky’s seminal work in the 1950s, the standard explanation for why only humans have language is that only human minds possess an innate “language faculty” in which the universal rules of human grammar have been encoded (e.g., Jackendoff, 2002; Pinker, 1994). The notion of a specialized language faculty tuned to a universal grammar fits nicely with the presumption that, outside of language, the rest of the human mind is more or less like that of any other ape. At the extreme, Hauser et al. (2002), for example, suggest that the sole qualitative difference between human and nonhuman cognition might be the computational mechanism of recursion putatively at the core of all human languages.

But there are about 5,000 to 8,000 distinct languages spoken by humans today. And as comparative linguists are increasingly pointing out, the remarkable diversity of all these languages makes the notion of a universal grammar difficult to sustain (Evans & Levinson, 2010). There are some languages with no derivational morphology or constituent structure; there are some that do not employ fixed orders of elements. Some languages have no words for numerals. Some have no sounds (e.g., sign languages). There are languages without adverbs or adjectives. And even recursion turns out to be a feature that has a thousand variations and is apparently absent

in the Amazonian language Pirahá (Everett, 2005). Nevertheless, despite this spectacular diversity, any normal human child can learn to comprehend and employ any human language. How is this possible?

Perhaps the answer lies in the fact that human languages have been shaped by the idiosyncrasies of the human mind, rather than the other way around (see Christiansen & Chater, 2008). In other words, the statistical properties of human languages evolved, by natural and cultural selection, to be just those that our brains were capable of learning. And the evidence suggests that many of aspects of language acquisition, such as word learning and grammar comprehension, rely on general-purpose systems—such as our ability to attribute beliefs to others and to make analogical inferences—that originally evolved for other purposes and still perform these nonlinguistic functions to this day (e.g., Bloom, 2000a; Christiansen & Kirby, 2003; Tomasello, 2000).

It is likely that evolution of human language profoundly “rewired” the human brain, as Bermudez (2005) and Bickerton (2009) have suggested. But there must have been a complex, coevolutionary relationship between the human cognitive architecture and human language in order for language to take root (see again Christiansen & Chater, 2008). And the comparative evidence suggests that the human brain must have been uniquely human before even the simplest proto-language could even get started (cf. Bickerton, 2009). So language alone cannot be the first, or even the most fundamental, difference between humans and nonhuman minds.

It's Not by Culture Alone

In their book, *Not by Genes Alone*, Richerson and Boyd (2004) demonstrate that the cultural transmission of beliefs, practices, and specialized knowledge has played a unique role in shaping the human mind. The dynamics of cultural evolution, they show, are analogous to those of biological evolution but nevertheless limited to the human species. Many animals may have “cultures” of their own, and young animals routinely learn from their elders (Hunt & Gray, 2004; Rendell & Whitehead, 2001; van Schaik et al., 2003; Whiten, 2000; Whiten et al., 1999). But nonhuman species never accumulate complex behaviors or knowledge over successive generations as human cultures do (Richerson & Boyd, 2004; Richerson, Boyd, & Henrich, 2002; Tennie, Call, & Tomasello, 2009). Is human culture, then, the source of human cognitive uniqueness?

Humans, it is true, are inordinately dependent on each other and exceptionally social. We are not just a “prosocial” animal; we are, as Gazzaniga (2009) put it, “the party animal.” Hrdy (2000, 2009) highlights the crucial role that cooperative maternal care has played in the evolution of our species’ peculiar cognitive abilities. The complexity and importance of cooperative interactions among humans is unrivaled in any other species (Burkart, Hrdy, & Van Schaik, 2009; Silk, 2003; Silk et al., 2005). And it is obvious that all of our species’ most salient achievements—from fire building to Twitter—would be unthinkable without the transmission of enormous bodies of knowledge between generations, the ability to assimilate symbolic cultural norms, and an inherently prosocial orientation. At the extreme, some researchers have argued that a human child kept alive on a deserted island and magically raised to adulthood without any human contact would “not differ very much—perhaps a little, but not very much” from other great apes (Tomasello & Rakoczy, 2003, p. 121).

But a close examination of the comparative data does not support the hypothesis that cultural learning and prosocial motivations alone make us unique. Many chimpanzees have been raised by devoted researchers under conditions that rival those of coddled human children. Yet after decades of enculturation, even the most gifted chimp’s ability to reason about the world does not rival that of the average human toddler (David Premack, & Premack, 2003; Penn, Holyoak, & Povinelli, 2008). Even more fundamentally, while the parents of many species often facilitate social learning among their offspring (Bender, Herzing, & Bjorklund, 2009; de Waal, 2001; Hoppitt et al., 2008), only human teachers take into account what their students do and do not know.

For example, one of the most widely publicized examples of “teaching” among animals comes from the stars of Animal Planet’s *Meerkat Manor* (Thornton, 2008; Thornton & McAuliffe, 2006). For meerkats, scorpions are a delicacy. But a scorpion’s stinger would be a serious threat to a young meerkat. So meerkat parents kill or disable the scorpions before giving them to their pups. At first, parents give new pups dead scorpions. Then, when the pups are a little older, the parents give them live scorpions with the stinger removed. Eventually, the young meerkats graduate to the whole package, stinger and all.

At first glance, this seems like a compelling example of human-like teaching—and the popular media has certainly portrayed meerkats in this way. But the same researchers who first documented the

meerkats' lesson plan were also the first to discover that meerkats do not actually cater their lessons to the skill level of the pup but to the quality of their begging calls (see also Madden, Kunc, English, & Clutton-Brock, 2009). Play a recording of a young pup and the mother will give her older pups a dead scorpion. Play a recording of an older pup and a meerkat mother will give her infant a live scorpion, stinger and all, even though this may kill her child.

Without the scaffolding provided by human culture, our species' cognitive achievements might be paltry indeed. But what remains to be explained is why only human children can learn the lessons that human parents have to teach and why only human teachers take into account what their students do and do not know.

It's Not Only Our "Theory of Mind"

Premack and Woodruff (1978) originally coined the term "Theory of Mind" (ToM) to refer to our ability to impute mental states to others—such as goals, intentions, beliefs, and doubts—and to use these hypothetical and unobservable entities to predict and explain others' behavior. This cognitive system properly counts as a "theory," Premack and Woodruff argued, because "such [mental] states are not directly observable, and the system can be used to make predictions about the behavior of others" (p. 515).

Whether any nonhuman animal has a ToM is a question that has been "fraught with controversy" from the get-go (Shettleworth, 2010). Over 10 years ago, Heyes (1998) reviewed the experimental work to date and concluded that there had been no "substantial progress" in answering this question. It is far from clear that any more progress has been made since then. Over the last decade, there has been a flurry of new experiments claiming to provide evidence that nonhuman animals understand at least some psychological states in others (see Call & Tomasello, 2008 for a review). But these claims remain controversial. Penn and Povinelli (2007b), for example, argue that there is still no evidence that nonhuman animals have anything even remotely resembling a ToM. As Emery and Clayton (2009) recently put it, whether one believes that animals reason about others' mental states or not may be largely a matter of "faith."

It is clear that social vertebrates of many taxa are adept at interpreting the observable *behavior* of others, including how conspecifics are likely to behave given those specific individuals' past behavior and the behavior of other conspecifics under similar circumstances

(Emery & Clayton, 2009; Povinelli & Vonk, 2004). And it is clear that many animals—including apes, dogs, monkeys, and corvids—are quite good at acting *as if* they were taking others' perspective into account in certain species-specific contexts (Call, Brauer, Kaminski, & Tomasello, 2003; Emery & Clayton, 2008; Flombaum & Santos, 2005; Kaminski, Brauer, Call, & Tomasello, 2009; Santos, Nissen, & Ferrugia, 2006). But nonhuman animals consistently fail any task that requires them to reason about another individual's beliefs when that individual's beliefs are different from their own (Call & Tomasello, 2008; Kaminski, Call, & Tomasello, 2008).

So does something about our social intelligence and our "prosocial" personalities explain what makes human cognition unique? Many researchers believe so (see chapters in Kappeler & Silk, 2010). Tomasello and his colleagues, for example, argue that humans have evolved a suite of unique skills for understanding other individuals as cooperative agents with whom one can share emotions, experience, and collaborative actions (Tomasello, 1999, 2008; Tomasello, Carpenter, Call, Behne, & Moll, 2005a, 2005b; Tomasello & Rakoczy, 2003). And Tomasello argues that this species-specific ability to participate with others in collaborative activities with shared goals and intentions (i.e., what he calls "shared intentionality") lies at the heart of human cognitive uniqueness.

Although nonhuman animals clearly do not have the ability to collaborate or communicate with each other in human-like ways, the hypothesis that our capacity for "shared intentionality" alone lies at the heart of the gap between human and nonhuman minds seems difficult to sustain. It is very hard to see how a discontinuity in social-cognitive abilities alone could explain the profound differences between human and nonhuman animals' abilities in other nonsocial domains, such as our ability to create complex tools, diagnose causal relationships and construct theories about the world (Penn & Povinelli, 2007a; Povinelli, 2000, *in press*). For example, chimpanzees and corvids are quite adept at using sticks to retrieve out-of-reach objects, but there are consistent and glaring gaps in their understanding of the causal relationships involved (Holzhaider, Hunt, Campbell, & Gray, 2008; Povinelli, 2000; Seed, Tebbich, Emery, & Clayton, 2006; Taylor, Hunt, Medina, & Gray, 2009; Visalberghi, 2000; Visalberghi & Tomasello, 1998).

Tomasello and his colleagues recently carried out a massive experiment to test their "cultural

intelligence” hypothesis by comparing the social and physical reasoning skills of 106 chimpanzees and 32 orangutans against those of 105 2–3-year-old human children (Herrmann, Call, Hernandez-Lloreda, Hare, & Tomasello, 2007). All subjects were exposed to a common battery of 16 tasks. One set of tasks tested subjects’ ability to keep track of the number and location of objects as they were moved around, out of sight, beneath opaque cups. Another set of tasks tested subjects’ ability to reason about causal-logical relationships. For example, the experimenters hid a reward inside one of two cups and then shuffled the cups so the subjects couldn’t tell which cup contained the reward. On some trials, the experimenters shook the empty cup, which made no noise. On other trials, the experimenters shook the baited cup, which made a loud rattling sound. The subjects had to infer which cup contained the reward without looking inside. A third set of tasks tested subjects’ social intelligence. For example, the experimenters demonstrated the correct solution to a problem or looked at where a reward was hidden and the subjects were graded on how well they responded to the experimenters’ cues.

Herrmann et al. (2007) reported that there was no significant difference between children’s and chimpanzees’ performance on the tests of perceptual, quantitative, and causal-logical reasoning. However, human children performed significantly better than chimpanzees and orangutans on the social tasks. “Human children are not overall more intelligent than other primates,” the authors concluded, “but instead have specialized skills for social cognition. They learn in a way that chimpanzees don’t learn.” (p. 1360)

But a close inspection of Herrmann et al.’s (2007) data tells a very different story. To claim that there was no significant difference between human and nonhuman subjects on tests in the physical domain, the authors lumped all the scores on the spatial, quantitative, and causal-logical tasks into one composite average. When the scores on the individual tasks are considered separately, it turns out that the only test of causal-logical reasoning on which chimpanzees outperformed children was a task in which subjects had to retrieve an out-of-reach reward with a wooden stick: that is, the very tool-use task that is the chimpanzee’s special forte. And on this critical task, the children were given 1 minute to retrieve the reward and the apes were given 2 minutes. On every other test of causal-logical reasoning, the children performed significantly better than both the chimpanzees and the orangutans. Given these

results, Herrmann et al. (2007, p. 1365) themselves admit that “it is possible that what is distinctively human is not social-cultural cognition as a specialized domain…Rather, what may be distinctive is the ability to understand unobserved causal forces in general, including (as a special case) the mental states of others as causes of behavior.”

The hypothesis that our prosocial tendencies alone lie at the heart of the gap between human and nonhuman minds also does little to explain why only humans are capable of learning to manipulate symbols in a rule-governed fashion. Symbol-trained animals have no trouble learning to use symbols to get what they want. But as we discussed in the section, “It’s Not Our ‘Language Instinct,’” symbol-trained animals never master even the most rudimentary grammatical rules. Learning a human language appears to require a suite of inferential abilities that are both domain general and uniquely human (Christiansen & Chater, 2008; Christiansen & Kirby, 2003). Indeed, in a different context, Tomasello himself has argued that human language learners rely on cognitive capacities—such as analogical reasoning and abstract rule learning—that are independent from ToM and absent in nonhuman animals (Tomasello, 2000). So while our ability and desire to participate in collaborative activities and take each others’ beliefs into account is surely a distinctive feature of the human mind, it cannot be the only or even the most basic one.

The Human Mind Is Not a Beak

The latest explanation for the gap between human and nonhuman minds is inspired by the neo-Darwinian hypothesis that minds are collections of specialized “modules,” each one evolved to fulfill a particular adaptive function in our ancestral niche (e.g., Barkow, Cosmides, & Tobby, 1992; Carruthers, 2006; Pinker, 1997). According to this hypothesis, human minds have a “language” instinct and scrub jays have a “nut-caching” instinct; and the differences between human and nonhuman minds are like the differences between the beaks of Darwin’s finches: multifarious, incremental, and highly specialized.

There is considerable merit to this hypothesis. Many aspects of animals’ minds clearly evolved for particular tasks: The spatial memory of food-caching birds is a particularly well-studied and compelling example of “adaptive specialization” (e.g., de Kort, Tebbich, Dally, Emery, & Clayton, 2006; Shettleworth, 2003). But the comparative evidence suggests that an animal’s intelligence is not exclusively modular or domain

specific. For example, animals who regularly use tools in the wild—such as capuchin monkeys and New Caledonian crows—are not necessarily more adept at solving novel tool-use problems than other closely related species who do not normally use tools—such as vervet monkeys and rooks (Bird & Emery, 2009; Santos, Pearson, Spaepen, Tsao, & Hauser, 2006; Tebbich, Seed, Emery, & Clayton, 2007). Furthermore, the same species that are the most adept at solving new tool-use problems in captivity also tend to be the most adept at solving novel problems in every other domain, from social interactions to symbol manipulation (Clayton & Emery, 2005; Emery & Clayton, 2004; Seed et al., 2009).

The same is true for humans. Human language is the product of innumerable adaptive specializations, ranging from the mechanics of vocal production to the reorganization of our temporal lobes (Pinker & Jackendoff, 2005). And we know that there are specific regions and systems in the human brain tuned to thinking about mental states (Saxe, 2006; Saxe & Powell, 2006) and manipulating complex tools (Johnson-Frey, 2004), to mention only a few examples of the many evolved specializations peculiar to our species. The human mind is clearly not the product of one lucky mutation (Dennett, 1995). But the massively modular hypothesis favored by neo-Darwinian psychologists does not explain why there is an analogous discontinuity between human and nonhuman minds across every domain of cognition, from spatial navigation and tool use to language and cooperation. If there were some animals that used fire, some that built baskets, and some that had language and culture, the “massively modular” hypothesis might be more credible. But we’re the only ones who do any of these things. So it seems highly implausible that the discontinuities in all of these domains are the result of independent, domain-specific “modules” or that the human mind is simply the serendipitous result of a large number of unrelated adaptations. It seems much more likely (not to mention parsimonious) that some common set of adaptations coevolved with these domain-specific adaptations to enable a qualitative leap in human intelligence across the board.

The crux of the matter then is to identify the core set of changes that enabled this qualitative leap forward in the human species and no other.

The Relational Reinterpretation Hypothesis

Here’s one possible hypothesis: One of the essential features of any animal’s mind is its ability to

learn about novel relations in an adaptive fashion. Honeybees navigate by remembering and recalculating the relation between landmarks and their hive (Menzel & Giurfa, 2006). Baboons make inductive generalizations about how to behave based on the changing relationships between conspecifics, and they respond differently to different *kinds* of tertiary relations (Bergman, Beehner, Cheney, & Seyfarth, 2003; Seyfarth & Cheney, 2003). Teach a rat that it gets a reward when it presses a lever and it will quickly grasp the goal-directed relation between its action and the outcome (Dickinson & Balleine, 2000); and rats don’t respond to a *causal* relationship in the same way as they do to a mere association (Blaisdell, Sawa, Leising, & Waldmann, 2006). Relations, in short, are the currency of cognition. But Penn et al. (2008) argue that humans take relations one huge step further than the rest.

Many animals can learn the instrumental relation between a symbol and an object in the world. But only human children learn to grasp the higher order relation between symbols in terms of abstract grammatical rules. Animals of many taxa can learn to use tools to solve specific tasks. But only human children develop theories to explain how the world works. Figuring out how others are likely to act based on their observable behavior is a cognitive feat that nearly every social species has mastered. Figuring out how others are likely to act based on what they *believe* is something only human children learn.

All these disparate examples of uniquely human cognition—from ToM and theory building to language and tool use—have something critical in common: They all involve relations that aren’t manifest in the perceptual features of the objects themselves: You can’t see a grammatical rule, smell a theory, or touch somebody else’s thoughts. These unobservable entities are all invented by the human mind (Vonk & Povinelli, 2006). And a large and growing body of research shows that the ability to represent and reason about the unobservable relation between perceptual relations and the abstract roles that entities play in those relations—that is, what researchers call *role-based relational reasoning* (see Holyoak, Chapter 13)—subserves all these various aspects of human cognition from childhood on (Gentner, 2003; Gentner, Holyoak, & Kokinov, 2001; Goodwin & Johnson-Laird, 2005; Halford, 1993; Halford, Wilson, & Phillips, 1998, 2010; Holyoak & Thagard, 1995; Johnson-Laird, 1983; and see Doumas & Hummel, Chapter 5). Arguably, role-based relational reasoning lies at the core of

what makes human cognition unique (Gentner, 2003; Halford et al., 2010; Holyoak & Thagard, 1995; Penn et al., 2008).

Role-based relational reasoning requires the capacity to represent structured relations in terms of the bindings between concrete, observable entities and functional, unobservable roles (see Holyoak, Chapter 13). According to the relational reinterpretation hypothesis (Penn et al., 2008), because nonhuman animals are incapable of forming role-based representations, they are limited to reasoning about the relationship between things they can see, touch, hear, or smell. Only human minds reinterpret perceptual relations in terms of higher order relations that can't be perceived, such as rules, analogies, stories, and theories. The available evidence suggests that all these different kinds of higher order relations are processed by a general-purpose, multi-component system that has been grafted onto the systems for reasoning about perceptual relationships we inherited from our nonhuman ancestors. Adding this extra level of relations to a mind is not like moving from a one-story to a two-story house. It's like moving from a two-dimensional to a three-dimensional world.

Although we are far from possessing a complete understanding of the neural bases of higher order role-based relational reasoning (see Holyoak, Chapter 13 for a review), the available evidence provides promising support for the relational reinterpretation hypothesis. It is now clear that the human brain contains a specialized system for learning and reasoning about higher order relationships that crucially involves the anterior-most region of the prefrontal cortex traditionally identified as Brodmann Area 10 (Buckner, Andrews-Hanna, & Schacter, 2008; Bunge & Zelazo, 2006; Christoff et al., 2001; Gilbert et al., 2006; Kroger et al., 2002; e.g., Miller, 2000; Ramnani & Owen, 2004; Wendelken, Nakhabenko, Donohue, Carter, & Bunge, 2008; and see Morrison & Knowlton, Chapter 6). The anterior prefrontal cortex has been implicated in every form of relational reasoning that is uniquely human: from planning for the future and making logical deductions to constructing causal explanations and empathizing with others' emotional states. Within the anterior prefrontal cortex, the more medial regions seem to play an instrumental role in empathizing with other people's feelings, pondering our own thoughts, planning for the future, and imagining hypothetical scenarios; and the more lateral regions seem implicated in reasoning about analogies, logical

relations, and abstract rules (Burgess, Dumontheil, & Gilbert, 2007; Gilbert et al., 2006; Halford et al., 2010; Wendelken et al., 2008). Not coincidentally, the anterior prefrontal cortex is one of the newest parts of the human brain and a part of our brain that differs dramatically from that of other apes (Preuss, 2000; Semendeferi et al., 2001). Many features of the human brain have changed since our lineage split from that of other apes, but arguably the changes in our anterior prefrontal cortex are the ones that lie behind our unique ability to reason about role-based relations and thus to master grammatically structured languages, reason about others' thoughts, manufacture wheels, and ponder the peculiar nature of our own minds.

Conclusions and Future Directions

Although the relational reinterpretation hypothesis may provide a promising framework for understanding what makes human cognition unique, it is far from complete and there are still many questions left unanswered. For example, although there is some evidence suggesting that the development of a ToM is causally dependent on higher order relational reasoning (Andrews, Halford, Bunch, Bowden, & Jones, 2003; Apperly & Butterfill, 2009; Zelazo, Jacques, Burack, & Frye, 2002), there has been no conclusive evidence showing that deficits in role-based relational reasoning must necessarily produce deficits in ToM functioning as the relational reinterpretation hypothesis claims. The same criticism applies to the relationship between language learning and higher order relations: Although there is good evidence that language comprehension and production involve higher order relational computations similar to those involved in analogical reasoning and ToM (see Christiansen & Kirby, 2003; Christoff et al., 2001; Halford et al., 2010; Waltz et al., 2004), we do not yet know to what extent role-based relational reasoning is necessary to support normal language use and development.

Worse, while neuroscientists have been relatively successful at identifying *where* in the brain higher order relationships are processed, we still know very little about *how* the human brain performs its remarkable computations or what those computations actually consist of. A few researchers have attempted to address this question in a neurally plausible fashion (Doumas & Hummel, 2005, and Chapter 5; e.g., Hummel & Holyoak, 2003, 2005). While we applaud these efforts and believe they have already shed new light on what makes human

cognition unique, our species' best models to date are still a very long way from accounting for the unique and still enigmatic features of the human mind.

References

- Amati, D., & Shallice, T. (2007). On the emergence of modern humans. *Cognition*, 103(3), 358–385.
- Andrews, G., Halford, G. S., Bunch, K. M., Bowden, D., & Jones, T. (2003). Theory of mind and relational complexity. *Child Development*, 74(5), 1476–1499.
- Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, 116(3), 953–970.
- Baldo, J. V., Dronkers, N. F., Wilkins, D., Ludy, C., Raskin, P., & Kim, J. (2005). Is problem solving dependent on language? *Brain and Language*, 92(3), 240–250.
- Barkow, J. H., Cosmides, L., & Tobby, J. (1992). *The Adapted mind: evolutionary psychology and the generation of culture*. Oxford University Press.
- Bender, C. E., Herzing, D. L., & Bjorklund, D. F. (2009). Evidence of teaching in Atlantic spotted dolphins (*Stenella frontalis*) by mother dolphins foraging in the presence of their calves. *Animal Cognition*, 12(1), 43–53.
- Bergman, T. J., Beehner, J. C., Cheney, D. L., & Seyfarth, R. M. (2003). Hierarchical classification by Rank and Kinship in baboons. *Science*, 302(5648), 1234–1236.
- Bermudez, J. L. (2003). *Thinking without words*. New York: Oxford University Press.
- Bermudez, J. L. (2005). *Philosophy of Psychology: A contemporary introduction*. London: Routledge.
- Bickerton, D. (2009). *Adam's tongue: How humans made language, how language made humans*. New York: Hill and Wang.
- Bird, C. D., & Emery, N. J. (2009). Insightful problem solving and creative tool modification by captive nontool-using rooks. *Proceedings of the National Academy of Sciences USA*, 106(25), 10370–10375.
- Blaisdell, A. P., Sawa, K., Leising, K. J., & Waldmann, M. R. (2006). Causal reasoning in rats. *Science*, 311(5763), 1020–1022.
- Bloom, P. (2000a). *How children learn the meaning of words*. Cambridge, MA: MIT Press.
- Bloom, P. (2000b). Language and thought: does grammar makes us smart? *Current Biology*, 10(14), R516-R517.
- Bloom, P. (2004). Can a dog learn a word? *Science*, 304, 1605–1606.
- Bloom, P., & Wynn, K. (1997). Linguistic cues in the acquisition of number words. *Journal of Child Language*, 24, 511–533.
- Bluff, L. A., Weir, A. A. S., Rutz, C., Wimpenny, J. H., & Kacelnik, A. (2007). Tool-related cognition in New Caledonian crows. *Comparative Cognition and Behavior Reviews*, 2, 1–25.
- Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: Anatomy, function, and relevance to disease. *Annals of the New York Academy of Science*, 1124, 1–38.
- Bunge, S., & Zelazo, P. (2006). A brain-based account of the development of rule use in childhood. *Current Directions in Psychological Science*, 15(3), 118–221.
- Burgess, P. W., Dumontheil, I., & Gilbert, S. J. (2007). The gateway hypothesis of rostral prefrontal cortex (area 10) function. *Trends in Cognitive Sciences*, 11(7), 290–298.
- Burkart, J. M., Hrdy, S. B., & Van Schaik, C. P. (2009). Cooperative breeding and human cognitive evolution. [Review]. *Evolutionary Anthropology*, 18(5), 175–186.
- Byrne, R. W. (1993). Do larger brains mean greater intelligence? *Behavioral and Brain Sciences*, 16, 696–697.
- Byrne, R. W. (1994). The evolution of intelligence. In P. J. B. Slater & T. R. Halliday (Eds.), *Behaviour and evolution* (pp. 223–265). Cambridge University Press.
- Byrne, R. W. (1996). Relating brain size to intelligence in primates. In P. A. Mellars & K. R. Gibson (Eds.), *Modeling the early human mind* (pp. 49–56). Cambridge, England: Macdonald Institute for Archaeological Research.
- Byrne, R. W., & Whiten, A. (1988). *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans*. New York: Clarendon Press.
- Call, J., Brauer, J., Kaminski, J., & Tomasello, M. (2003). Domestic dogs (*Canis familiaris*) are sensitive to the attentional state of humans. *Journal of Comparative Psychology*, 117(3), 257–263.
- Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, 12(5), 187–192.
- Carruthers, P. (2002). The cognitive functions of language. *Behavioral and Brain Sciences*, 25(6), 657–726.
- Carruthers, P. (2006). *The architecture of the mind*. Oxford University Press.
- Chimpanzee Sequencing and Analysis Consortium. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055), 69–87.
- Cho, S., Moody, T. D., Fernandino, L., Poldrack, R. A., Cannon, T. D., Knowlton, B. J. et al. (2010). Neural mechanisms of fluid intellectual processes: Event-related fMRI identifies prefrontal loci for relational integration and interference resolution.
- Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, 31(05), 489–509.
- Christiansen, M. H., & Kirby, S. (2003). Language evolution: Consensus and controversies. *Trends in Cognitive Sciences*, 7(7), 300–307.
- Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J. K., Holyoak, K. J., & Gabriel, J. D. (2001). Rostrolateral prefrontal cortex involvement in relational integration during reasoning. *Neuroimage*, 14(5), 1136–1149.
- Clark, A. (2001). Reasons, robots and the extended mind. *Mind and Language*, 16(2), 121–145.
- Clark, A. (2006). Language, embodiment, and the cognitive niche. *Trends in Cognitive Sciences*, 10(8), 370–374.
- Clayton, N., & Emery, N. (2005). Corvid cognition. *Current Biology*, 15(3), R80–R81.
- Darwin, C. (1871). *The descent of man, and selection in relation to sex*. London: John Murray.
- de Kort, S. R., Tebbich, S., Dally, J. M., Emery, N. J., & Clayton, N. S. (2006). The comparative cognition of caching. In T. R. Zentall & E. A. Wasserman (Eds.), *Comparative cognition* (pp. 602–618). New York, NY: Oxford University Press.
- de Waal, F. B. M. (2001). *The ape and the sushi master: Cultural reflections by a primatologist*. New York: Basic Books.
- Deacon, T. W. (1997). *The symbolic species*. New York: W. W. Norton.
- Dehaene, S. (1997). *The number sense*. New York: Oxford University Press.
- Dennett, D. C. (1995). *Darwin's dangerous idea: evolution and the meanings of life*. New York: Simon & Schuster.
- Dennett, D. C. (1996). *Kinds of minds: Toward an understanding of consciousness*. New York: Basic Books.
- Diamond, J. M. (1992). *The third chimpanzee: The evolution and future of the human animal*. New York: HarperCollins.

- Dickinson, A., & Balleine, B. (2000). Causal cognition and goal-directed action. In C. M. Heyes & L. Huber (Eds.), *The evolution of cognition* (pp. 185–204). Cambridge, MA: MIT Press.
- Doumas, L., & Hummel, J. E. (2005). Approaches to modeling human mental representations: What works, what doesn't and why. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 73–91). New York: Cambridge University Press.
- Dunbar, R. I. M. (1993). Co-evolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences*, 16, 681–735.
- Dunbar, R. I. M. (1998). Neocortex size predicts group size in carnivores and some insectivores. *Ethology*, 104, 695–708.
- Dunbar, R. I. M., & Shultz, S. (2007). Evolution in the social brain. *Science*, 317(5843), 1344–1347.
- Dupre, J. (2003). *Darwin's legacy: What evolution means today*. New York, NY: Oxford University Press.
- Emery, N. J., & Clayton, N. S. (2004). The mentality of crows: Convergent evolution of intelligence in corvids and apes. *Science*, 306(5703), 1903–1907.
- Emery, N. J., & Clayton, N. S. (2008). How to build a scrub-jay that reads minds. In S. Itakura & K. Fujita (Eds.), *Origins of the social minds: Evolutionary and developmental views* (pp. 65–97). Hicom, Japan: Springer.
- Emery, N. J., & Clayton, N. S. (2009). Comparative social cognition. *Annual Review of Psychology*, 60(1), 87–113.
- Evans, N., & Levinson, S. C. (2010). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5), 429–492.
- Everett, D. L. (2005). Cultural constraints on grammar and cognition in Piraha: Another look at the design features of human language. *Current Anthropology*, 46, 621–646.
- Finarelli, J. A., & Flynn, J. J. (2007). The evolution of encephalization in caniform carnivorans. *Evolution*, 61(7), 1758–1772.
- Finarelli, J. A., & Flynn, J. J. (2009). Brain-size evolution and sociality in Carnivora. *Proceedings of the National Academy of Sciences USA*, 106, 9345–9349.
- Flombaum, J. I., & Santos, L. R. (2005). Rhesus monkeys attribute perceptions to others. *Current Biology*, 15(5), 447–452.
- Garfield, J. L., Peterson, C. C., & Perry, T. (2001). Social cognition, language acquisition and the development of the theory of mind. *Mind and Language*, 16(5), 494–541.
- Gazzaniga, M. (2000). Cerebral specialization and the interhemispheric communication: Does the corpus callosum enable the human condition? *Brain*, 123(Pt. 7), 1293–1326.
- Gazzaniga, M. (2008). *Human: The science behind what makes us unique*. New York: Ecco.
- Gazzaniga, M. (2009). Humans: The party animal. *Daedalus*, 138(3), 21–34.
- Gentner, D. (2003). Why we're so smart. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 195–235). Cambridge, MA: MIT Press.
- Gentner, D., & Christie, S. (2008). Relational language supports relational cognition in humans and apes. *Behavioral and Brain Sciences*, 31(02), 136–137.
- Gentner, D., Holyoak, K. J., & Kokinov, B. N. (Eds.). (2001). *The analogical mind: Perspectives from cognitive science*. Cambridge, MA: MIT Press.
- Gentner, D., & Kurtz, K. J. (2005). Learning and using relational categories. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. W. Wolff (Eds.), *Categorization inside and outside the lab* (pp. 151–175). Washington, DC: American Psychological Association.
- Gilbert, S. J., Spengler, S., Simons, J. S., Steele, J. D., Lawrie, S. M., Frith, C. D., & Burgess, P. W. (2006). Functional specialization within rostral prefrontal cortex (Area 10): A meta-analysis. *Journal of Cognitive Neuroscience*, 18(6), 932–948.
- Goldin-Meadow, S. (2003). *The resilience of language: What gesture creation in deaf children can tell us about how all children learn language*. New York: Psychology Press.
- Goodwin, G. P., & Johnson-Laird, P. (2005). Reasoning about relations. *Psychological Review*, 112(2), 468–493.
- Grosenick, L., Clement, T. S., & Fernald, R. D. (2007). Fish can infer social rank by observation alone. *Nature*, 445(7126), 429–432.
- Halford, G. S. (1993). *Children's understanding: The development of mental models*. Hillsdale, NJ: Erlbaum.
- Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, 21(6), 803–831; discussion 831–864.
- Halford, G. S., Wilson, W. H., & Phillips, S. (2010). Relational knowledge: The foundation of higher cognition. *Trends in Cognitive Sciences*, 14(11), 497–505.
- Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it and how did it evolve? *Science*, 298(5598), 1569–1579.
- Herman, L. M. (1986). Cognition and language competencies of bottlenosed dolphins. In R. J. Schusterman, J. H. Thomas, & F. G. Wood (Eds.), *Dolphin cognition and behavior: A comparative approach* (pp. 221–251). Hillsdale, NJ: Erlbaum.
- Herman, L. M., Richards, D. G., & Wolz, J. P. (1984). Comprehension of sentences by bottlenosed dolphins. *Cognition*, 16, 129–219.
- Herrmann, E., Call, J., Hernandez-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science*, 317(5843), 1360–1366.
- Heyes, C. M. (1998). Theory of mind in nonhuman primates. *Behavioral and Brain Sciences*, 21(1), 101–114; discussion 115–148.
- Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought*. Cambridge, MA: MIT Press.
- Holzhaider, J. C., Hunt, G. R., Campbell, V. M., & Gray, R. D. (2008). Do wild New Caledonian crows (*Corvus monedulaoides*) attend to the functional properties of their tools? *Animal Cognition*, 11(2), 243–254.
- Hoppitt, W. J., Brown, G. R., Kendal, R., Rendell, L., Thornton, A., Webster, M. M., & Laland, K. N. (2008). Lessons from animal teaching. *Trends in Ecology and Evolution*, 23, 486–493.
- Hrdy, S. B. (2000). *Mother nature: Maternal instinct and how they shape the human species*. New York: Ballantine Books.
- Hrdy, S. B. (2009). *Mothers and others: The evolutionary origins of mutual understanding*. Cambridge, MA: Belknap Press.
- Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, 110, 220–264.
- Hummel, J. E., & Holyoak, K. J. (2005). Relational reasoning in a neurally plausible cognitive architecture. *Current Directions in Psychological Science*, 14(3), 153–157.
- Hunt, G. R., & Gray, R. D. (2004). The crafting of hook tools by wild New Caledonian crows. *Proceedings of the Royal Society B: Biological Sciences*, 271(0), S88–S90.

- Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. New York: Oxford University Press.
- Johnson-Frey, S. H. (2004). The neural bases of complex tool use in humans. *Trends in Cognitive Sciences*, 8(2), 71–78.
- Johnson-Laird, P. (1983). *Mental models*. Cambridge, England: Cambridge University Press.
- Kako, E. (1999). Elements of syntax in the systems of three language-trained animals. *Animal Learning and Behavior*, 27, 1–14.
- Kaminski, J., Brauer, J., Call, J., & Tomasello, M. (2009). Domestic dogs are sensitive to a human's perspective. *Behaviour*, 146, 979–998.
- Kaminski, J., Call, J., & Fischer, J. (2004). Word learning in a domestic dog: Evidence for "fast mapping." *Science*, 304, 1682–1683.
- Kaminski, J., Call, J., & Tomasello, M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition*, 109(2), 224–234.
- Kappeler, P., & Silk, J. (Eds.). (2010). *Mind the gap: Tracing the origins of human universals*. New York, NY: Springer.
- Kroger, J. K., Saab, F. W., Fales, C. L., Bookheimer, S. Y., Cohen, M. S., & Holyoak, K. J. (2002). Recruitment of anterior dorsolateral prefrontal cortex in human reasoning: A parametric study of relational complexity. *Cerebral Cortex*, 12, 477–485.
- Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive Psychology*, 50(4), 315–353.
- Madden, J. R., Kunc, H.-J. P., English, S., & Clutton-Brock, T. H. (2009). Why do meerkat pups stop begging? *Animal Behaviour*, 78(1), 85–89.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: W. H. Freeman.
- McGrew, W. C. (1992). *Chimpanzee material culture: Implications for human evolution*. New York, NY: Cambridge University Press.
- Menzel, R., & Giurfa, M. (2006). Dimensions of cognition in an insect, the honeybee. *Behavioral and Cognitive, and Neuroscience Review*, 5(1), 24–40.
- Miller, E. K. (2000). The prefrontal cortex and cognitive control. *Nature Reviews Neuroscience*, 1(1), 59–65.
- Mithen, S. (1996). *The prehistory of the mind*. Thames, England: Hudson.
- Mithen, S. (2006). *The singing Neanderthals: The origins of music, language, mind and body*. : Nicolson.
- Peake, T. M., Terry, A. M., McGregor, P. K., & Dabelsteen, T. (2002). Do great tits assess rivals by combining direct experience with information gathered by eavesdropping? *Proceedings of the Royal Society B: Biological Sciences*, 269(1503), 1925–1929.
- Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, 31(2), 109–178.
- Penn, D. C., & Povinelli, D. J. (2007a). Causal cognition in human and nonhuman animals: A comparative, critical Review. *Annual Review of Psychology*, 58, 97–118.
- Penn, D. C., & Povinelli, D. J. (2007b). On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362, 731–744.
- Pepperberg, I. M. (2002). *The Alex studies: Cognitive and communicative abilities of grey parrots*. Cambridge, MA: Harvard University Press.
- Peterson, C. C., & Siegal, M. (1995). Deafness, conversation and the theory of mind. *Journal of Child Psychology, Child Psychiatry and Allied Disciplines*, 36, 459–474.
- Peterson, C. C., & Siegal, M. (2000). Insights into theory of mind from deafness and autism. *Mind and Language*, 15(1), 123–145.
- Pinker, S. (1994). *The language instinct*. New York: W. Morrow.
- Pinker, S. (1997). *How the mind works*. New York, NY: W.W. Norton.
- Pinker, S., & Jackendoff, R. (2005). The faculty of language: What's special about it? *Cognition*, 95, 201–236.
- Povinelli, D. J. (2000). *Folk physics for apes: The chimpanzee's theory of how the world works*. Oxford, England: Oxford University Press.
- Povinelli, D. J. (2012). *World without Weight: Perspectives on an alien mind*. Oxford University Press.
- Povinelli, D. J., & Vonk, J. (2004). We don't need a microscope to explore the chimpanzee's mind. *Mind and Language*, 19(1), 1–28.
- Premack, D. (2004). Is language the key to human intelligence? *Science*, 303(5656), 318–320.
- Premack, D., & Premack, A. (2003). *Original intelligence: Unlocking the mystery of who we are*. New York: McGraw-Hill.
- Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 4, 515–526.
- Preuss, T. M. (2000). What's human about the human brain. In M. S. Gazzaniga (Ed.), *The new cognitive neurosciences* (pp. xiv, 1419). Cambridge, MA: MIT Press.
- Preuss, T. M. (2004). What is it like to be a human? In M. Gazzaniga (Ed.), *The cognitive neurosciences* (3rd ed., pp. 5–22). Cambridge, MA: MIT Press.
- Ramnani, N., & Owen, A. M. (2004). Anterior prefrontal cortex: Insights into function from anatomy and neuroimaging. *Nature Reviews Neuroscience*, 5(3), 184–194.
- Rendell, L., & Whitehead, H. (2001). Culture in whales and dolphins. *Behavioral and Brain Sciences*, 24, 309–382.
- Richerson, P., Boyd, R., & Henrich, J. (2002). The cultural evolution of human cooperation. In P. Hammerstein (Ed.), *The genetic and cultural evolution of cooperation* (pp 357–388). Cambridge, MA: MIT Press.
- Richerson, P. J., & Boyd, R. (2004). *Not by genes alone: How culture transformed human evolution*. Chicago, IL: University Of Chicago Press.
- Rightmire, G. P. (2004). Brain size and encephalization in early to mid-Pleistocene Homo. *American Journal of Physical Anthropology*, 124(2), 109–123.
- Rilling, J. K., & Seligman, R. A. (2002). A quantitative morphometric comparative analysis of the primate temporal lobe. *Journal of Human Evolution*, 42(5), 505–533.
- Roth, G., & Dicke, U. (2005). Evolution of the brain and intelligence. *Trends in Cognitive Sciences*, 9(5), 250–257.
- Rushton, J. P., & Ankney, C. D. (2009). Whole brain size and general mental ability: A review. *International Journal of Neuroscience*, 119(5), 691–731.
- Sandler, W., Meir, I., Padden, C., & Aronoff, M. (2005). From the cover: The emergence of grammar: Systematic structure in a new language. *Proceedings of the National Academy of Sciences USA*, 102(7), 2661–2665.

- Santos, L. R., Nissen, A. G., & Ferrugia, J. A. (2006). Rhesus monkeys, *Macaca mulatta*, know what others can and cannot hear. *Animal Behaviour*, 71(5), 1175–1181.
- Santos, L. R., Pearson, H., Spaepen, G., Tsao, F., & Hauser, M. (2006). Probing the limits of tool competence: Experiments with two non-tool-using species (*Cercopithecus aethiops* and *Saguinus oedipus*). *Animal Cognition*, 9(2), 94–109.
- Savage-Rumbaugh, S., Shanker, S. G., & Taylor, T. J. (1998). *Apes, language, and the human mind*. New York: Oxford University Press.
- Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology*, 16(2), 235–239.
- Saxe, R., & Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science*, 17(8), 692–699.
- Schoenemann, P. T. (2006). Evolution of the size and functional areas of the human brain. *Annual Review of Anthropology*, 35(1), 379–406.
- Schoenemann, P. T., Budinger, T. F., Sarich, V. M., & Wang, W. S. (2000). Brain size does not predict general cognitive ability within families. *Proceedings of National Academy of Sciences USA*, 97(9), 4932–4937.
- Schoenemann, P. T., Sheehan, M. J., & Glotzer, L. D. (2005). Prefrontal white matter volume is disproportionately larger in humans than in other primates. *Nature Neuroscience*, 8(2), 242–252.
- Schusterman, R. J., & Gisiner, R. (1989). Please parse the sentence: Animal cognition in the procustean bed of linguistics. *Psychological Record*, 39, 3–18.
- Schusterman, R. J., & Krieger, K. (1986). Artificial language comprehension and size transportation by a California sea lion (*Zalophus Californianus*). *Journal of Comparative Psychology*, 100, 348–355.
- Seed, A. M., Emery, N. J., & Clayton, N. S. (2009). Intelligence in corvids and apes: A case of convergent evolution? *Ethology*, 115, 401–420.
- Seed, A. M., Tebbich, S., Emery, N. J., & Clayton, N. S. (2006). Investigating physical cognition in rooks (*Corvus frugilegus*). *Current Biology*, 16, 697–701.
- Semendeferi, K., Armstrong, E., Schleicher, A., Zilles, K., & Van Hoesen, G. W. (2001). Prefrontal cortex in humans and apes: A comparative study of area 10. *American Journal of Physical Anthropology*, 114(3), 224–241.
- Senghas, A., Kita, S., & Ozyurek, A. (2004). Children creating core properties of language: Evidence from an emerging sign language in Nicaragua. *Science*, 305(5691), 1779–1782.
- Seyfarth, R. M., & Cheney, D. L. (2003). The structure of social knowledge in monkeys. In F. B. M. de Waal & P. L. Tyack (Eds.), *Animal social complexity: Intelligence, culture and individualized societies* (pp. 230–248). Cambridge, MA: Harvard University Press.
- Shettleworth, S. J. (2003). Memory and hippocampal specialization in food-storing birds: Challenges for research on comparative cognition. *Brain, Behavior and Evolution*, 62(2), 108–116.
- Shettleworth, S. J. (2010). *Cognition, evolution and behavior*. New York, NY: Oxford University Press.
- Siegal, M., Varley, R. A., & Want, S. C. (2001). Mind over grammar: Reasoning in aphasia and development. *Trends in Cognitive Sciences*, 5(7), 296–301.
- Silk, J. B. (2003). The evolution of cooperation in primate groups. In H. Gintis, S. Bowles, R. Boyd, & E. Fehr (Eds.), *Moral sentiments and material interests: On the foundations of cooperation in economic life* (pp. 43–73). Cambridge, MA: MIT Press.
- Silk, J. B., Brosnan, S. F., Vonk, J., Henrich, J., Povinelli, D. J., Richardson, A. S.,... Schapiro, S. J. (2005). Chimpanzees are indifferent to the welfare of unrelated group members. *Nature*, 437(7063), 1357–1359.
- Striedter, G. (2005). *Principles of brain evolution*. Sunderland, MA: Sinauer Associates.
- Taylor, A. H., Hunt, G. R., Medina, F. S., & Gray, R. D. (2009). Do New Caledonian crows solve physical problems through causal reasoning? *Proceedings of the Royal Society B: Biological Sciences*, 276(1655), 247–254.
- Tebbich, S., Seed, A. M., Emery, N. J., & Clayton, N. S. (2007). Non-tool-using rooks (*Corvus frugilegus*) solve the trap-tube task. *Animal Cognition*, 10(2), 225–231.
- Tennie, C., Call, J., & Tomasello, M. (2009). Ratcheting up the ratchet: On the evolution of cumulative culture. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1528), 2405–2415.
- Thornton, A. (2008). Variation in contributions to teaching by meerkats. *Proceedings of the Royal Society B: Biological Sciences*, 275(1644), 1745–1751.
- Thornton, A., & McAuliffe, K. (2006). Teaching in wild meerkats. *Science*, 313(5784), 227–229.
- Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.
- Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition*, 74(3), 209–253.
- Tomasello, M. (2008). *The origins of human communication*. Cambridge, MA: MIT Press.
- Tomasello, M., & Call, J. (1997). *Primate cognition*. New York: Oxford University Press.
- Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005a). In search of the uniquely human. *Behavioral and Brain Sciences*, 28(5), 721–727.
- Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005b). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28, 675–691.
- Tomasello, M., & Rakoczy, H. (2003). What makes human cognition unique? From individual to shared to collective intentionality. *Mind and Language*, 18(2), 121–147.
- Uhlhaas, P. J., & Singer, W. (2006). Neural synchrony in brain disorders: Relevance for cognitive dysfunctions and pathophysiology. *Neuron*, 52(1), 155–168.
- van Schaik, C. P., Ancrenaz, M., Borgen, G., Galdikas, B., Knott, C. D., Singleton, I., et al. (2003). Orangutan cultures and the evolution of material culture. *Science*, 299(5603), 102–105.
- Varela, F., Lachaux, J. P., Rodriguez, E., & Martinerie, J. (2001). The brainweb: Phase synchronization and large-scale integration. *Nature Reviews Neuroscience*, 2(4), 229–239.
- Varki, A., & Altheide, T. K. (2005). Comparing the human and chimpanzee genomes: Searching for needles in a haystack. *Genome Research*, 15(12), 1746–1758.
- Varley, R. A., Klessinger, N. J., Romanowski, C. A., & Siegal, M. (2005). Agrammatic but numerate. *Proceedings of the National Academy of Sciences USA*, 102(9), 3519–3524.
- Varley, R. A., & Siegal, M. (2000). Evidence for cognition without grammar from causal reasoning and 'theory of mind' in an agrammatic aphasic patient. *Current Biology*, 10(12), 723–726.

- Varley, R. A., Siegal, M., & Want, S. C. (2001). Severe impairment in grammar does not preclude theory of mind. *Neurocase*, 7(6), 489–493.
- Visalberghi, E. (2000). Tool use behaviour and the understanding of causality in primates. In E. Thommen & H. Kilcher (Eds.), *Comparer ou prédire: Exemples de recherches en psychologie comparative aujourd’hui* [Compare or predict: examples of research in comparative psychology today] (pp. 17–35). Fribourg, Switzerland: Les Editions Universitaires.
- Visalberghi, E., & Tomasello, M. (1998). Primate causal understanding in the physical and psychological domains. *Behavioural Processes*, 42(2–3), 189–203.
- Vonk, J., & Povinelli, D. J. (2006). Similarity and difference in the conceptual systems of primates: The unobservability hypothesis. In T. Zentall & E. A. Wasserman (Eds.), *Comparative cognition: Experimental explorations of animal intelligence* (pp. 363–387). Oxford, England: Oxford University Press.
- Waltz, J. A., Knowlton, B. J., Holyoak, K. J., Boone, K. B., Back-Madruga, C., McPherson, S.,...Miller, B. L. (2004). Relational integration and executive function in Alzheimer’s disease. *Neuropsychology*, 18(2), 296–305.
- Wendelken, C., Nakhabenko, D., Donohue, S. E., Carter, C. S., & Bunge, S. A. (2008). “Brain is to thought as stomach is to ??”: Investigating the role of rostral-lateral prefrontal cortex in relational reasoning. *Journal of Cognitive Neuroscience*, 20(4), 682–693.
- Whiten, A. (2000). Primate culture and social learning. *Cognitive Science*, 24(3), 477–508.
- Whiten, A., Goodall, J., McGrew, W. C., Nishida, T., Reynolds, V., Sugiyama, Y., et al. (1999). Cultures in chimpanzees. *Nature*, 399(6737), 682–685.
- Zelazo, P. D., Jacques, S., Burack, J. A., & Frye, D. (2002). The relation between theory of mind and rule use: Evidence from persons with autism-spectrum disorders. *Infant and Child Development*, 11, 171–195.

New Perspectives on Language and Thought

Lila Gleitman and Anna Papafragou

Abstract

In this chapter we discuss the question of whether, how, where, and to what extent language plays a causally fundamental role in creating categories of thought, and in organizing and channeling thought that is already mentally present. In general, both logic and currently available evidence suggest a disclaimatory view of strongest proposals (Benjamin Whorf, 1956) according to which particulars of certain human languages are important progenitors of thought, such that elements of perception or conception would be permanently altered by learning one or another language. However, several credible lines of experimental and developmental evidence suggest significant influence of linguistic representation during online processing in many cognitive and perceptual domains: Insofar as languages differ in the short-term processing demands that they pose to listeners, interpretational outcomes and styles, including characteristic ambiguity resolution, may look quite different cross-linguistically as a function of concomitant population differences (e.g., age-group) and task demands.

Key Words: categorical perception, Whorf, linguistic relativity, linguistic determinism

The presence of language is one of the central features that distinguish humans from other species. Even in very early infancy, during the (misnamed) prelinguistic stage of life, infants respond positively to strangers who are speaking in the special melodies of the exposure language, but they shrink away from those speaking a different language or dialect (Kinzler, Shutts, DeJesus, & Spelke, 2009). Cultures materially define themselves by the way that they and “others” speak, down to the smallest details. Blood has in many times and places been spilled in consequence. A famous case, and the origin of the word itself, is the biblical tale of *shibboleth*.¹

In light of this intimate bond between language and cultural identification, it is easy to understand the intense interest for laypersons and specialists alike on the topic of this chapter: *the relations between language and thought*. Many people actually identify these two notions; they share the intuition that they

think “in” language, hence that the absence of language would, *ipso facto*, be the absence of thought itself. One compelling version of this self-reflection is Helen Keller’s (1955) report that her recognition of the signed symbol for “water” triggered thought and emotional processes that had theretofore—and consequently—been utterly absent. Statements to the same or related effect come from the most diverse intellectual sources: “The limits of my language are the limits of my world” (Wittgenstein, 1922/1961); and “The fact of the matter is that the ‘real world’ is to a large extent unconsciously built upon the language habits of the group” (Sapir, 1941, as cited in Whorf, 1956, p. 75). On this kind of supposition, we may have no way to think many thoughts, conceptualize many of our ideas, without language, or outside of and independent of language. Moreover, different communities of humans, speaking different languages, would think differently to just the extent

that their languages differ from one another. But is this so? Could it be so? That depends on how we unpack the notions so far alluded to so informally.

Do We Think “in” Language?

In the obvious sense, language has powerful and specific effects on thought. After all, that’s what it is for, or at least that is one of the things it is for: to transfer ideas from one mind to another mind. Imagine Eve telling Adam *Apples taste great*. This fragment of linguistic information, as we know, caused Adam to entertain a new thought with profound effects on his world knowledge, inferencing, and subsequent behavior. Much of human communication is an intentional attempt to modify the thoughts and attitudes of others in just this way. This information transmission function is crucial for the structure and survival of cultures and societies in all their known forms (also see Rai, Chapter 29).

Traditionally, language has been considered mainly in this conduit role, as the vehicle for the expression of thought rather than as its progenitor. From Noam Chomsky’s universalist perspective, for example, the forms and contents of all particular languages derive, in large part, from an antecedently specified cognitive substance and architecture, and therefore they provide a rich diagnostic of human conceptual commonalities:

Language is a mirror of mind in a deep and significant sense. It is a product of human intelligence... By studying the properties of natural languages, their structure, organization, and use, we may hope to learn something about human nature; something significant, if it is true that human cognitive capacity is the truly distinctive and most remarkable characteristic of the species. (Chomsky, 1975, p. 4)

This view is not proprietary to the rationalist position for which Chomsky is speaking here. Classical empiricist thought maintained that our concepts (sensory discriminations aside) derive from experience with properties, things, and events in the world and not, originally, from language:

To give a child an idea of scarlet or orange, of sweet or bitter, I present the objects, or in other words, convey to him these impressions; but proceed not so absurdly, as to endeavor to produce the impressions by exciting the ideas. (Hume, 1739/2000; Book I)

And as a consequence of such experience with *things*, *ideas* arise in the mind and can receive linguistic labels:

If we will observe how children learn languages, we shall find that, to make them understand what the names of simple ideas or substances for, people ordinarily *show them the thing whereof they would have them have the idea*; and then repeat to them the name that stands for it... (Locke, 1690/1964, Book 3.IX.9; italics ours)

Our question in this chapter is how far and in what ways this chain may operate in reverse, such that language causes thought to be what it is. The issues here were raised most forcefully in the writings of Benjamin Whorf and Eric Sapir in the first half of the 20th century.² According to Whorf, the grammatical and lexical resources of individual languages heavily constrain the conceptual representations available to their speakers.

We are thus introduced to a new principle of relativity, which holds that all observers are not led by the same physical evidence to the same picture of the universe, unless their linguistic backgrounds are similar, or can in some way be calibrated. (Whorf, 1956, p. 214)

This linguistic-relativistic view, in its richest form, entails that linguistic categories will be the “program and guide for an individual’s mental activity” (ibid, p. 212), including categorization, memory, reasoning, and decision making. If this is right, then the study of different linguistic systems may throw light onto the diverse modes of thinking encouraged or imposed by such systems. The importance of this position cannot be overestimated: Language here becomes a vehicle for the growth of *new* concepts—those which were not theretofore in the mind, and perhaps could not have been there without the intercession of linguistic experience. Thus, it poses a challenge to the venerable view that one could not acquire a concept that one could not antecedently entertain (Plato, 5–4th century BCE; Descartes, 1662; Fodor, 1975, *inter alia*). At the limit it is a proposal for how new thoughts can arise in the mind as a result of experience with language rather than as a result of experience with the world of objects and events.

By the 1950s the Whorf-Sapir hypothesis began to percolate into psychological theorizing, and it seemed to proponents to provide a route to understanding how cognitive categories formed and jelled in the developing human mind. A major figure in this history was Roger Brown, the great social and developmental psychologist who framed much of the field of language acquisition in the modern era. Brown (1957) performed a simple and elegant experiment

that demonstrated an effect of lexical categorization on the inferred meaning of a new word. Young children were shown a picture, for example, of hands that seemed to be kneading confetti-like stuff in an overflowing bowl. Some children were told, *Show me the sib*. They pointed to the bowl (a solid rigid object). Others were told, *Show me some sib*. They pointed to the confetti (an undifferentiated mass of stuff). Others were told, *Show me sibbing*. They pointed to the hands and made kneading motions with their own hands (an action or event). Plainly, the same stimulus object was represented differently depending on the linguistic cues to the lexical categories count noun, mass noun, and verb. That is, the lexical categories themselves have notional correlates, at least in the minds of these young English speakers.

Some commentators have argued that the kinds of cues exemplified here, for example, that persons, places, and things surface as nouns, are universal and thus can play causal roles in the acquisition of language by learners who are predisposed to find just these kinds of syntactic-semantic correlations "natural" (e.g., Baker, 2001; Bloom, 1994a; Fisher, 1996; Gleitman, 1990; Landau & Gleitman, 1985; Lidz, Gleitman & Gleitman, 2003; Pinker, 1984). Brown saw his result the other way around. He supposed that languages would vary arbitrarily in these form mappings onto conceptual categories. Those world properties thus yoked together by language would cause a (previously uncommitted) infant learner to conceive of them as meaningfully related in some ways.

In learning a language, therefore, it must be useful to discover the semantic correlates for the various parts of speech; for this discovery enables the learner to use the part-of-speech membership of a new word as a first cue to its meaning... Since [grammatical categories] are strikingly different in unrelated languages, the speakers [of these languages] may have quite different cognitive categories. (Brown, 1957, p. 5)

These ideas have continued to be explored in the cognitive literature ever since. One recent formulation states:

Instead of language merely reflecting the cognitive development which permits and constrains its acquisition, language is thought of as potentially catalytic and transformative of cognition.
(Bowerman & Levinson, 2001a, p. 13)

In the strongest interpretations, the categories of language essentially become the default categories of thought:

We surmise that language structure... provides the individual with a system of representation, some isomorphic version of which becomes highly available for incorporation as a default conceptual representation. Far more than developing simple habituation, use of the linguistic system, we suggest, actually forces the speaker to make computations he or she might otherwise not make. (Pederson et al., 1998, p. 586)

Before turning to the recent literature on language and thought, so conceived, we want to emphasize that most current contributors fall somewhere between the extremes of such views. To our knowledge, none of those who are currently advancing linguistic-relativistic themes and explanations believe that infants enter into language acquisition in a state of complete conceptual nakedness, later redressed (perhaps we should say "dressed") by linguistic information. Rather, infants are believed to possess some "core knowledge" that enters into the first categorizations of objects, properties, and events in the world (e.g., Baillargeon, 1993; Carey, 1982, 2008; Gelman & Spelke, 1981; Gibson & Spelke, 1983; Kellman, 1996; Leslie & Keeble, 1987; Mandler, 1996; Prasada, Ferenz, & Haskell, 2002; Quinn, 2003; Spelke, Breinlinger, Macomber, & Jacobson, 1992). The viable question is how richly specified this innate basis may be and how experience refines, enhances, and transforms the mind's original furnishings; and, finally, whether specific language knowledge may be one of these formative or transformative aspects of experience. To our knowledge, none of those who adopt a nativist position on these matters reject as a matter of a priori conviction the possibility that there could be effects of language on thought. For instance, some particular natural language might formally mark a category that another does not; two languages might draw a category boundary at different places; and two languages might differ in the computational resources they require to make manifest a particular distinction or category. These differences might, in turn, influence the representation or processing machinery for speech and comprehension.

We will try to draw out aspects of these issues within several domains in which commentators and investigators are currently trying to disentangle cause and effect in the interaction of language and thought. We cannot discuss it all, of course, or even very much of what is currently in print on this topic. There is too much of it (for recent anthologies,

see Bowerman & Levinson, 2001a; Gentner & Goldin-Meadow, 2003; Gumperz & Levinson, 1996; Malt & Wolff, 2010).

Border Wars: Where Does Language End and Inference Begin?

We begin with a very simple question: Do our thoughts actually take place in a specific natural language? If so, it would immediately follow that Whorf was right all along, since speakers of Korean and Spanish, or Swahili and Hopi would have to think systematically different thoughts.

There are several reasons to suppose that, if tenable at all, such a position needs to be reined in considerably. This is because, if language directly expresses our thought, it seems to make a poor job of it. Consider, for example, this sentence from the preceding section:

1. There is too much of it.

Leaving aside, for now, the problems of anaphoric reference (what is “it”?), the sentence still has at least two interpretations that are compatible with its discourse context:

- 1a. ‘There is too much written on linguistic relativity to fit into this article.’
- 1b. ‘There is too much written on linguistic relativity.’ (*Period!*)

We authors had one of these two interpretations in mind (guess which one). We had a thought and expressed it as (1) but English failed to render that thought unambiguously, leaving things open as between (1a) and (1b). One way to think about what this example portends is that language just cannot, or in practice does not, express all and only what we mean. Rather, language use offers hints and guideposts to hearers, such that they can usually reconstruct what the speaker had in mind by applying to the uttered words a good dose of common sense, *aka* thoughts, inferences, and plausibilities in the world.

The question of just how to apportion the territory between the underlying semantics of sentences and the pragmatic interpretation of the sentential semantics is, of course, far from settled in linguistic and philosophical theorizing. Consider the sentence *It is raining*. Does this sentence directly—that is, as an interpretive consequence of the linguistic representation itself—convey an assertion about rain falling *here*? That is, *in the immediate geographical environment of the speaker*? Or does the sentence itself—the linguistic representation—convey only

that rain is falling, leaving it for the common sense of the listener to deduce that the speaker likely meant raining here and now rather than raining today in Bombay or on Mars; likely too that if the sentence was uttered indoors, the speaker more likely meant *here* to convey “just outside of here” than “right here, as the roof is leaking.”³ The exact division of labor between linguistic semantics and pragmatics has implications for the language-thought issue, since the richer (one claims that) the linguistic semantics is, the more likely it is that language guides our mental life. Without going into detail, we will argue that linguistic semantics cannot fully envelop and substitute for inferential interpretation; hence, the representations that populate our mental life cannot be identical to the representations that encode linguistic (semantic) meaning.

Language Is Sketchy; Thought Is Rich

There are several further reasons to believe that thought processes are not definable over representations that are isomorphic to linguistic representations. One is the pervasive ambiguity of words and sentences. *Bat*, *bank*, and *bug* all have multiple meanings in English and hence are associated with multiple concepts, but these concepts themselves are clearly distinct in thought, as shown *inter alia* by the fact that one may consciously construct a pun. Moreover, several linguistic expressions including pronouns (*he*, *she*) and indexicals (*here*, *now*) crucially rely on context for their interpretation while the thoughts they are used to express are usually more specific. Our words are often semantically general, that is, they fail to make distinctions that are nevertheless present in thought: *uncle* in English does not semantically specify whether the individual comes from the mother’s or the father’s side, or whether he is a relative by blood or marriage, but usually the speaker who utters this word (*my uncle...*) possesses the relevant information. Indeed, lexical items typically take on different interpretations tuned to the occasion of use (*He has a square face*; *The room is hot*) and depend on inference for their precise construal in different contexts (e.g., the implied action itself is systematically different when we *open an encyclopedia**a can**an umbrella**a book* or when an instance of that class of actions is performed to serve different purposes: *open the window to let in the evening breeze**the cat*). Moreover, there are cases where linguistic output does not even encode a complete thought/proposition (*Tomorrow*, *Maybe*). Finally, the presence of implicatures and other kinds of pragmatic

inference ensures that—to steal a line from the Mad Hatter—while speakers generally mean what they say, they do not and could not say exactly what they mean.

From this and related evidence, it appears that linguistic representations underdetermine the conceptual contents they are used to convey: Language is *sketchy* compared to the richness of our thoughts (for related discussions, see Fisher & Gleitman, 2002; Papafragou, 2007). In light of the limitations of language, time, and sheer patience, language users make reference by whatever catch-as-catch-can methods they find handy, including the waitress who famously told another that “The ham sandwich wants his check” (Nunberg, 1978). In this context, *Table 8, the ham sandwich and the man seated at Table 8* are communicatively equivalent. What chiefly matters to talkers and listeners is that successful reference be made, whatever the means at hand. If one tried to say all and exactly what one meant, conversation could not happen; speakers would be lost in thought. Instead, conversation involves a constant negotiation in which participants estimate and update each others’ background knowledge as a basis for what needs to be said versus what is mutually known and inferable (e.g., Bloom, 2000; Clark, 1992; Grice, 1975; Sperber & Wilson, 1986).

In limiting cases, competent listeners ignore linguistically encoded meaning if it patently differs from (their estimate of) what the speaker intended, for instance, by smoothly and rapidly repairing slips of the tongue. Oxford undergraduates had the wit, if not the grace, to snicker when Reverend Spooner said, or is reputed to have said, “Work is the curse of the drinking classes.” Often the misspeaking is not even consciously noticed but is repaired to fit the thought, evidence enough that the word and the thought are two different matters.⁴ The same latitude for thought to range beyond established linguistic means holds for the speakers, too. Wherever the local linguistic devices and locutions seem insufficient or overly constraining, speakers invent or borrow words from another language, devise similes and metaphors, and sometimes make permanent additions and subtractions to the received tongue. It would be hard to understand how they do so if language were itself, and all at once, both the format and vehicle of thought.

Arbitrary and Inconsistent Encodings

The cases just mentioned refer to particular tokenings of meanings in the idiosyncratic interactions between people. A related problem arises

when languages categorize aspects of the world in ways that are complex and inconsistent. An example is reported by Malt, Sloman, Gennari, Shi, and Wang (1999). They examined the vocabulary used by English, Spanish, and Chinese subjects to label the various containers we bring home from the grocery store full of milk, juice, ice cream, bleach, or medicine (e.g., *jugs, bottles, cartons, boxes*). As the authors point out, containers share names based not only on some perceptual resemblances but also on very local and particular conditions, with size, contents, shape, substance, nature of the contents, not to speak of the commercial interests of the purveyor, all playing interacting and shifting roles. For instance, in present-day American English, a certain plastic container that looks like a bear with a straw stuck in its head is called “a juice box,” though it is not boxy either in shape (square or rectangular) or typical constitution (your prototypical American box is made of cardboard). The languages Malt et al. studied differ markedly in the set of terms available for this domain and also in how their subjects extended these terms to describe diverse new containers. Speakers of the three languages differed in which objects (old and new) they classified together by name. For example, a set of objects distributed across the sets of *jugs, containers, and jars* by English speakers were unified by the single label *frasco* by Spanish speakers. Within and across languages not everything square is a box, not everything glass is a bottle, not everything *not* glass is *not* a bottle, and so on. The naming, in short, is a complex mix resulting from perceptual resemblances, historical influences, and a generous dollop of arbitrariness. Yet Malt et al.’s subjects did not differ much (if at all) from each other in their classification of these containers by overall similarity rather than by name. Nor were the English and Spanish, as one might guess, more closely aligned than, say, the Chinese and Spanish. So here we have a case where cross-linguistic practice groups objects in a domain in multiple ways that have only flimsy and sporadic correlations with perception, without discernible effect on the non-linguistic classificatory behaviors of users.⁵

So far we have emphasized that language is a relatively impoverished and underspecified vehicle of expression, which relies heavily on inferential processes outside the linguistic system for reconstructing the richness and specificity of thought. If correct, this seems to place rather stringent limitations on how language could serve as the original engine and sculptor of our conceptual life. Phrasal

paraphrase, metaphor and figurative language are heavily relied on to carry ideas that may not be conveniently lexicalized or grammaticalized. Interpretive flexibility sufficient to overcome these mismatches is dramatically manifested by simultaneous translators at the United Nations who more or less adequately convey the speakers' thoughts using the words and structures of dozens of distinct languages, thus crossing not only differences in the linguistic idiom but enormous gulfs of culture and disagreements in belief and perspective.

Despite these many disclaimers, it is possible to maintain the idea that certain formal properties of language causally affect thought in more local, but still important, ways. In the remainder of this chapter we consider two currently debated versions of the view that properties of language influence aspects of perception, thinking, and reasoning. The first is that language exerts its effects more or less *directly and permanently*, by revising either the mental categories, shifting the boundaries between them, or changing their prominence ("salience"). The second is that particulars of a language exert *indirect and transient* effects imposed during the rapid-fire business of talking and understanding. The present authors believe that the latter position, which we will explicate and expand as we go along, comes closer than the former to unifying the present experimental literature, and, in essence, reunites the Whorf-inspired literature with what we might call "ordinary psycholinguistics," the machinery of online comprehension.

Use It or Lose It: When Language Reorganizes the Categories of Thought

We begin with the most famous and compelling instance of language properties reconstructing perceptual categories: categorical perception of the phoneme (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Liberman, 1970; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Werker & Lalonde, 1988).

Children begin life with the capacity and inclination to discriminate among all of the acoustic-phonetic properties by which languages encode distinctions of meaning, a result famously documented by Peter Eimas (Eimas, Siqueland, Jusczyk, & Vigorito, 1971) using a dishabituation paradigm (for details and significant expansions of this basic result see, e.g., Jusczyk, 1985; Mehler & Nespor, 2004; Werker & DesJardins, 1995). These authors showed that an infant will work (e.g., turn its head

or suck on a nipple) to hear a syllable such as *ba*. After some period of time, the infant habituates; that is, its sucking rate decreases to some base level. The high sucking rate can be reinstated if the syllable is switched to, say, *pa*, demonstrating that the infant detects the difference. These effects are heavily influenced by linguistic experience. Infants only a year or so of age—just when true language is making its appearance—have become insensitive to phonetic distinctions that are not phonemic (play no role at higher levels of linguistic organization) in the exposure language (Werker & Tees, 1984). While these experience-driven effects are not totally irreversible in cases of long-term second-language immersion, they are pervasive and dramatic (for discussion, see Best, McRoberts, & Sithole, 1988; Werker & Logan, 1985). Without special training or unusual talent, the adult speaker-listener can effectively produce and discriminate the phonetic categories required in the native tongue, and little more. Not only that, these discriminations are categorical in the sense that sensitivity to within-category phonetic distinctions is poor and sensitivity at the phonemic boundaries is especially acute.

When considering these findings in the context of linguistic relativity, one might be tempted to write them off as a limited tweaking at the boundaries of acoustic distinctions built into the mammalian species, a not-so-startling sensitizing effect of language on perception (Aslin, 1981; Aslin & Pisoni, 1980). Moreover, it is unlikely that the limits on perception and production imposed by early learning are absolute. After all, depending on age, talent, and motivation, they can be altered once again in subsequent learning of a second (or third, etc.) language (Werker & Lalonde, 1988).

But a more radical restructuring, specific to particular languages, occurs as these phonetic elements are organized into higher level phonological categories. For example, American English speech regularly lengthens vowels in syllables ending with a voiced consonant (compare *ride* and *write*) and neutralizes the *t/d* distinction in favor of a dental flap in certain unstressed syllables. The effect is that (in most dialects) the consonant sounds in the middle of *rider* and *writer* are indistinguishable if removed from their surrounding phonetic context. Yet the English-speaking listener perceives a *d/t* difference in these words all the same, and—except when asked to reflect carefully—fails to notice the characteristic difference in vowel length that his or her own speech faithfully reflects. The complexity of

this phonological reorganization is often understood as a reconciliation (interface) of the cross-cutting phonetic and morphological categories of a particular language. *Ride* ends with a *d* sound; *write* ends with a *t* sound; morphologically speaking, *rider* and *writer* are just *ride* and *write* with *er* added on; therefore, the phonetic entity between the syllables in these two words must be *d* in the first case and *t* in the second. Morphology trumps phonetics (Bloch & Trager, 1942; Chomsky, 1964; for extensions to alphabetic writing, see Gleitman & Rozin, 1977).

Much of the literature on linguistic relativity can be understood as adducing related reconstructions in various perceptual and conceptual domains that are mapped onto language. Is it the case that distinctions of lexicon or grammar made regularly in one's language sensitize one to these distinctions and suppress or muffle others? Even to the extent of radically and permanently reorganizing the domain? We now look at some likely further cases.

The Perception of Hue

Languages differ in their terms for color just as they do in their phonetic and phonemic inventories. A number of factors favor color variables in the study of potential influences of language on thought. First, there is a powerful tradition of psychophysical measurement in this area that allows for the creation of test materials that can be scaled and quantitatively compared, at least roughly, for difference magnitudes, discriminability, and so on. Second, the fact that humans can discriminate hundreds of thousands, if not millions, of hues, coupled with the fact that it is impossible to learn a word for each, makes this domain a likely repository of linguistic difference. Third, the case of hue appears quite analogous to the well-studied instance of learning effects on phonetic categorization, thus a plausible immediate extension in the relevant regard.

Accordingly, a very large descriptive and experimental literature has been directed toward the question of whether color memory, learning, and similarity are influenced by color category-boundaries in the languages of the world. Significant evidence supports the view that color labeling is at least partly conditioned by universal properties of perception. Berlin and Kay (1969), in a cross-linguistic survey, showed that color vocabularies develop under strong universal constraints that are unlikely to be describable as effects of cultural diffusion (for recent discussion and amplifications, see especially

Regier, Kay, Gilbert, & Ivry, 2010). Nevertheless there is considerable variance in the number of color terms encoded, so it can be asked whether these linguistic labeling practices affect perception. Heider and Oliver (1972) made a strong case that they do not. They reported that the Dugum Dani, a preliterate Papuan tribe of New Guinea with only two color labels (roughly, warm-dark and cool-light), remembered and categorized new hues that they were shown in much the same way as English speakers who differ from them both culturally and linguistically.

Intriguing further evidence of the independence of perception and labeling practices comes from red-green color-blind individuals (deutanopes; Jameson & Hurvich, 1978). The perceptual similarity space of the hues for such individuals is systematically different from that of individuals with normal trichromatic vision. Yet a significant subpopulation of deutanopes *names* hues, even of new things, consensually with normal-sighted individuals and consensually orders these hue *labels* for similarity as well. That is, these individuals do not order a set of color chips by similarity with the reds at one end, the greens at the other end, and the oranges somewhere in between (rather, by alternating chips that the normal trichromat sees as reddish and greenish; that is what it means to be color blind). Yet they do organize the color words with *red* semantically at one end, *green* at the other, and *orange* somewhere in between. In the words of Jameson and Hurvich:

the language brain has learned denotative color language as best it can from the normal population of language users, exploiting whatever correlation it has available by way of a reduced, or impoverished, sensory system, whereas the visual brain behaves in accordance with the available sensory input, ignoring what its speaking counterpart has learned to say about what it sees. (1978, p. 154)

Contrasting findings had been reported earlier by Brown and Lenneberg (1954), who found that colors that have simple verbal labels are identified more quickly than complexly named ones in a visual search task (e.g., color chips called "blue" are, on average, found faster among a set of colors than chips called "purplish blue," etc.), suggesting that aspects of naming practices do influence recognition. In a series of recent studies in much the same spirit, Gilbert, Regier, Kay, and Ivry (2006; see also Regier, Kay, & Cook, 2005; Regier, Kay, & Khetarpal, 2009) have shown that reaction time in

visual search is longer for stimuli with the same label (e.g., two shades both called “green” in English) than for stimuli with different labels (one a consensual “blue” and one a consensual “green”). Crucially, however, this was the finding only when the visual stimuli were delivered to the right visual field (RVF), that is, projecting to the left, language-dominant, hemisphere. Moreover, the RVF advantage for differently labeled colors disappeared in the presence of a task that interferes with verbal processing but not in the presence of a task of comparable difficulty that does not disrupt verbal processing (see also Kay & Kempton, 1984; Winawer, Witthoft, Frank, Wu, & Boroditsky, 2007). This response style is a well-known index of categorical perception, closely resembling the classical results for phoneme perception.

Looking at the literature in broadest terms, then, and as Regier et al. (2010) discuss in an important review, the results at first glance seem contradictory: On the one hand, perceptual representations of hue reveal cross-linguistic labeling commonalities and are independent of such terminological differences as exist within these bounds. On the other hand, there are clear effects of labeling practices, especially in speeded tasks, where within-linguistic category responses are slower and less accurate than cross-category responses. The generalization appears to be that when language is specifically mobilized as a task requirement (e.g., the participant is asked for a verbal label) or when linguistically implicated areas of the brain are selectively measured, the outcomes are sensitive to linguistic categories; otherwise, less so or not at all: Language tasks recruit linguistic categories and functions that do not come into play in nonlinguistic versions of very similar tasks.⁶ As we next show, this generalization holds as well in a variety of further domains where linguistic effects on thinking have been explored.

Objects and Substances

The problem of reference to *stuff* versus *objects* has attracted considerable attention because it starkly displays the indeterminacy in how language refers to the world (Chomsky, 1957; Quine, 1960). Whenever we indicate some physical object, we necessarily indicate some portion of a substance as well; the reverse is also true. Languages differ in their expression of this distinction. Some languages make a grammatical distinction that roughly distinguishes object from substance (Chierchia, 1998; Lucy & Gaskins, 2001). Count nouns in such languages

denote individuated entities, for example, object kinds. These are marked in English with determiners like *a*, *the*, and are subject to counting and pluralization (*a horse*, *horses*, *two horses*). Mass nouns typically denote nonindividuated entities, for example, substance rather than object kinds. These are marked in English with a different set of determiners (*more porridge*), and they need an additional term that specifies quantity to be counted and pluralized (*a tube of toothpaste* rather than *a toothpaste*).

Soja, Carey, and Spelke (1991) asked whether children approach this aspect of language learning already equipped with the ontological distinction between things and substances, or whether they are led to make this distinction through learning count/mass syntax. Their subjects, English-speaking 2-year-olds, did not yet make these distinctions in their own speech. Soja et al. taught these children words in reference to various types of unfamiliar displays. Some were solid objects such as a T-shaped piece of wood, and others were nonsolid substances such as a pile of hand cream with sparkles in it. The children were shown such a sample, named with a term presented in a syntactically neutral frame that identified it neither as a count nor as a mass noun, for example, *This is my blicket* or *Do you see this blicket?* In extending these words to new displays, 2-year-olds honored the distinction between object and substance. When the sample was a hard-edged solid object, they extended the new word to all objects of the same shape, even when made of a different material. When the sample was a non-solid substance, they extended the word to other-shaped puddles of that same substance but not to shape matches made of different materials. Soja et al. took this finding as evidence of a conceptual distinction between objects and stuff, independent of and prior to the morphosyntactic distinction made in English.

This interpretation was put to stronger tests by extending such classificatory tasks to languages that differ from English in these regards: Either these languages do not grammaticalize the distinction, or they organize it in different ways (see Lucy, 1992; Lucy & Gaskins, 2001, for findings from Yucatec Mayan; Mazuka & Friedman, 2000; Imai & Gentner, 1997, for Japanese). Essentially, these languages’ nouns all start life as mass terms, requiring a special grammatical marker (called *a classifier*) if their quantity is to be counted. One might claim, then, that substance is in some sense linguistically basic for Japanese, whereas objecthood is basic for English speakers

because of the dominance of its count-noun morphology.⁷ So if children are led to differentiate object and substance reference by the language forms themselves, the resulting abstract semantic distinction should differ cross-linguistically. To test this notion, Imai and Gentner replicated Soja et al.'s original tests with Japanese and English children and adults. Some of their findings appear to strengthen the evidence for a universal prelinguistic ontology that permits us to think both about individual objects and about portions of stuff, for both American and Japanese children (even 2-year-olds) extended names for complex hard-edged nonsense objects on the basis of shape rather than substance. Thus, the lack of separate grammatical marking did not put the Japanese children at a disadvantage in this regard.

But another aspect of the results hints at a role for language itself in categorization. For one thing, the Japanese children tended to extend names for mushy hand-cream displays according to their substance, while the American children were at chance for these items. There were also discernible language effects on word extension for certain very simple stimuli (e.g., a kidney-bean-shaped piece of colored wax) that seemed to fall at the ontological midline between object and substance. While the Japanese at ages 2 and 4 were at chance on these items, the English speakers showed a tendency to extend words for them by shape.

How are we to interpret these results? Several authors have concluded that ontological boundaries literally shift to where language makes its cuts; that the substance/object distinction works much like the categorical perception effects we noticed for phonemes (and perhaps colors; see also Gentner & Boroditsky, 2001). Lucy and Gaskins (2001) bolstered this interpretation with evidence that populations speaking different languages differ increasingly with increasing age. While their young Mayan speakers are much like their English-speaking peers, by age 9 years members of the two communities differ significantly in relevant classificatory and memorial tasks. The implication is that long-term use of a language influences ontology, with growing conformance of concept grouping to linguistic grouping. Of course the claim is not for a rampant reorganization of thought; only for boundary shifting. Thus, for displays that clearly fall to one side or the other of the object/substance boundary, the speakers of all the tested languages sort the displays in the same ways.

It may be of some importance that suitable stimuli—those falling in the border area between stuff and thing—are hard to devise and instantiate, as we will discuss further. For this and related reasons, neither the findings nor the interpretations of such experiments are easy to come by. In one attempted replication, Mazuka and Friedman (2000) failed to reproduce Lucy's effects for Mayan- versus English-speaking subjects' classificatory performance for the predicted further case of Japanese. As these authors point out, the sameness in this regard of Japanese and English speakers, and the difference in this regard between Mayan and English speakers, suggests that obtained population differences may be more cultural and educational than linguistic.

In fact, there is another interpretation of these results that does not imply that language is altering the very categories of perception and thought. Rather, the result may again be limited to the influence of linguistic categories on linguistic performances, as we have noted before for the cases of phoneme and hue perception. This time the ultimate culprit is the necessarily sketchy character of most utterances, given ordinary exigencies of time and attention. One does not say (or rarely says), "Would you please set the table that is made of wood, is 6 feet in length, and is now standing in the dining room under the chandelier?" One says instead just enough to allow reference making to go through in a particular situational context. "Just enough," however, itself varies from language to language owing to differences in the basic vocabulary. Interpretations from this perspective have been offered by many commentators. Bowerman (1996), Brown (1957), Landau, Dessaegn, and Goldberg (2009), Landau and Gleitman (1985), Slobin (1996, 2001), and Papafragou, Massey, and Gleitman (2006), among others, propose that native speakers not only learn and use the individual lexical items their language offers but also learn the *kinds* of meanings typically expressed by a particular grammatical category in their language, and they come to expect new members of that category to have similar meanings. Languages differ strikingly in their most common forms and locutions—preferred fashions of speaking, to use Whorf's phrase. These probabilistic patterns could bias the interpretation of *new words*. Such effects come about in experiments when subjects are offered language input (usually nonsense words) under conditions in which implicitly known form-to-meaning patterns in the language might hint at how the new word is to be interpreted.

Let us reconsider the Imai and Gentner (1997) object-substance effects in light of this hypothesis. As we saw, when the displays themselves were of nonaccidental-looking hard-edged objects, subjects in both language groups opted for the object interpretation. But when the world was uninformative (e.g., for softish waxy lima bean shapes), the listeners fell back upon linguistic cues if available. No relevant morphosyntactic clues exist in Japanese, and so Japanese subjects chose at random for these indeterminate stimuli. For the English-speaking subjects, the linguistic stimulus too was in a formal sense interpretively neutral: *This blicket* is a template that accepts both mass and count nouns (*this horse/toothpaste*). But here principle and probability part company. Recent experimentation leaves no doubt that child and adult listeners incrementally exploit probabilistic facts about word use to guide the comprehension process online (e.g., Gleitman, January, Nappa, & Trueswell, 2007; Snedeker, Thorpe, & Trueswell, 2001; Tanenhaus, 2007; Trueswell, Sekerina, Hill, & Logrip, 1999). In the present case, any English speaker equipped with even a rough subjective probability counter should take into account the great preponderance of count nouns to mass nouns in English and so conclude that a new word *blicket*, used to refer to some indeterminate display, is very probably a new count noun rather than a new mass noun. Count nouns, in turn, tend to denote individuals rather than stuff and so have shape predictivity (Landau, Smith, & Jones, 1998; Smith, 2001). On this interpretation, it is not that speaking English leads one to tip the scales toward object representations of newly seen referents for perceptually ambiguous items; only that hearing English leads one to tip the scales toward count-noun representation of newly heard nominals in linguistically ambiguous structural environments. Derivatively, then, count syntax hints at object representation of the newly observed referent. Because Japanese does not have a corresponding linguistic cue, subjects choose randomly between the object/substance options where world-observation does not offer a solution. Such effects can be expected to increase with age as massive lexical-linguistic mental databases are built, consistent with the findings from Lucy and Gaskins (2001).⁸

Li, Dunham, and Carey (2009) recently tested the language-on-language interpretation conjectured by Fisher and Gleitman (2002) and Gleitman and Papafragou (2005), using an expanded set of object-like, substance-like, and neutral stimuli, in the Imai

and Gentner (1997) paradigm. They replicated the prior finding in several comparisons of Mandarin and English speakers. However, they added a new task, one that, crucially, did not require the subjects to interpret the meaning of the noun stimuli. This manipulation completely wiped out the cross-linguistic effect. As so often, the implication is that it is the linguistic nature of the task that elicits linguistic categories and functions. Languages differ in their vocabulary and structural patterns, impacting the procedures by which forms resolve to their meanings. But in nonlinguistic tasks, individuals with different linguistic backgrounds are found to respond in terms of the same conceptual categories.

Spatial Relationships

Choi and Bowerman (1991) studied the ways in which common motion verbs in Korean differ from their counterparts in English. First, Korean motion verbs often contain location or geometric information that is more typically specified by a spatial preposition in English. For example, to describe a scene in which a cassette tape is placed into its case, English speakers would say, "We put the tape *in the case*." Korean speakers typically use the verb *kkita* to express the *put in* relation for this scene. Second, *kkita* does not have the same extension as English *put in*. Both *put in* and *kkita* describe an act of putting an object in a location; but *put in* is used for all cases of containment (fruit in a bowl, flowers in a vase), while *kkita* is used only in case the outcome is a tight fit between two matching shapes (tape in its case, one Lego piece on another, glove on hand). Notice that there is a cross-classification here: While English appears to collapse across tightnesses of fit, Korean makes this distinction but conflates across *putting in* versus *putting on*, which English regularly differentiates. Very young learners of these two languages have already worked out the language-specific classification of such motion relations and events in their language, as shown by both their usage and their comprehension (Choi & Bowerman, 1991).

Do such cross-linguistic differences have implications for spatial cognition? McDonough, Choi, and Mandler (2003) focused on spatial contrasts between relations of tight containment versus loose support (grammaticalized in English by the prepositions *in* and *on* and in Korean by the verbs *kkita* and *nohta*) and tight versus loose containment (both grammaticalized as *in* in English but separately as *kkita* and *nehta* in Korean). They showed

that prelinguistic infants (9- to 14-month-olds) in both English- and Korean-speaking environments are sensitive to such contrasts, and so are Korean-speaking adults (see also Hespos & Spelke, 2000, who show that 5-month-olds are sensitive to this distinction). However, their English-speaking adult subjects showed sensitivity only to the tight containment versus loose support distinction, which is grammaticalized in English (*in* vs. *on*). The conclusion drawn from these results was that some spatial relations that are salient during the prelinguistic stage become less salient for adult speakers if their language does not systematically encode them: “Flexible infants become rigid adults.”

This interpretation again resembles the language-on-language effects in other domains, but in this case by no means as categorically as for the perception of phoneme contrasts. For one thing, the fact that English speakers learn and readily use verbs like *jam*, *pack*, and *wedge* weakens any claim that the lack of common terms seriously diminishes the availability of categorization in terms of tightness of fit. One possibility is that the observed language-specific effects with adults are due to verbal mediation: Unlike preverbal infants, adults may have turned the spatial classification task into a linguistic task. Therefore, it is useful to turn to studies that explicitly compare performance when subjects from each language group are instructed to classify objects or pictures by *name*, versus when they are instructed to classify the same objects by *similarity*.

In one such study, Li, Gleitman, Gleitman, and Landau, (1997) showed Korean- and English-speaking subjects pictures of events such as putting a suitcase on a table (an example of *on* in English, and of “loose support” in Korean). For half the subjects from each language group (each tested fully in their own language), these training stimuli were labeled by a videotaped cartoon character who performed the events (*I am Miss Picky and I only like to put things on things. See?*), and for the other subjects the stimuli were described more vaguely (...and I only like to do things like this. See?). Later categorization of new instances followed language in the labeling condition: English speakers identified new pictures showing tight fits (e.g., a cap put on a pen) as well as the original loose-fitting ones as belonging to the category that Miss Picky likes, but Korean speakers generalized only to new instances of loose fits. These language-driven differences radically diminished in the similarity sorting condition, in which the word (*on* or *nohta*) was not invoked; in this case the

categorization choices of the two language groups were essentially the same.

The “language-on-language” interpretation thus unifies the various laboratory effects in dealing with spatial relations, much as it does for hue perception, and for the object-substance distinction.

Motion

Talmy (1985) described two styles of motion expression that are typical for different languages: Some languages, including English, usually use a verb plus a separate path expression to describe motion events. In such languages, manner of motion is encoded in the main verb (e.g., *walk*, *crawl*, *slide*, or *float*), while path information appears in non-verbal elements such as particles, adverbials, or prepositional phrases (e.g., *away*, *through the forest*, *out of the room*). In Greek or Spanish, the dominant pattern instead is to include path information within the verb itself (e.g., Greek *bjeno* “exit” and *beno* “enter”); the manner of motion often goes unmentioned or appears in gerunds, prepositional phrases, or adverbials (*trehontas* “running”). These patterns are not absolute. Greek has motion verbs that express manner, and English has motion verbs that express path (*enter*, *exit*, *cross*). But several studies have shown that children and adults have learned these dominant patterns. Berman and Slobin (1994) showed that child and adult Spanish and English speakers vary in the terms that they most frequently use to describe the very same picture-book stories, with English speakers displaying greater frequency and diversity of manner of motion verbs. Papafragou, Massey, and Gleitman (2002) showed the same effects for the description of motion scenes by Greek- versus English-speaking children and, much more strongly, for Greek- versus English-speaking adults. Reasonably enough, the early hypothesis from Slobin and Berman was that the difference in language typologies of motion leads their speakers to different cognitive analyses of the scenes that they inspect. In the words of these authors, “children’s attention is heavily channeled in the direction of those semantic distinctions that are grammatically marked in the language” (Berman & Slobin, 1994), a potential salience or prominence effect of the categories of language onto the categories of thought.

Later findings did not sustain so strong a hypothesis, however. Papafragou, Massey, and Gleitman (2002) tested their English- and Greek- speaking subjects on either (*a*) memory of path or manner

details of motion scenes, or (*b*) categorization of motion events on the basis of path or manner similarities. Even though speakers of the two languages exhibited an asymmetry in encoding manner and path information in their verbal descriptions, they did not differ from each other in terms of classification or memory for path and manner.⁹ Similar results have been obtained for Spanish versus English by Gennari, Sloman, Malt, and Fitch (2002). Corroborating evidence also comes from studies by Munnich, Landau, and Dosher (2001), who compared English, Japanese, and Korean speakers' naming of spatial locations and their spatial memory for the same set of locations. They found that, even in aspects where languages differed (e.g., encoding spatial contact or support), there was no corresponding difference in memory performance across language groups.

Relatedly, the same set of studies suggests that the mental representation of motion and location is independent of linguistic naming *even within a single language*. Papafragou et al. (2002) divided their English- and Greek-speaking subjects' verbal descriptions of motion according to whether they included a path or manner verb, regardless of native language. Though English speakers usually chose manner verbs, sometimes they produced path verbs; the Greek speakers varied too but with the preponderances reversed. It was found that verb choice did not predict memory for path/manner aspects of motion scenes, or choice of path/manner as a basis for categorizing motion scenes. In the memory task, subjects who had used a path verb to describe a scene were no more likely to detect later path changes in that scene than subjects who had used a manner verb (and vice versa for manner). In the classification task, subjects were not more likely to name two motion events they had earlier categorized as most similar by using the same verb. Naming and cognition, then, are distinct under these conditions: Even for speakers of a single language, the linguistic resources mobilized for labeling underrepresent the cognitive resources mobilized for cognitive processing (e.g., memorizing, classifying, reasoning, etc.; see also Papafragou & Selimis, 2010b, for further evidence). An obvious conclusion from these studies of motion representation is that the conceptual organization of space and motion is robustly independent of language-specific labeling practices; nevertheless, specific language usage influences listeners' interpretation of the speaker's intended meaning if

the stimulus situation leaves such interpretation unresolved.¹⁰

Other recent studies have shown that motion event representation is independent of language even at the earliest moments of event apprehension. Papafragou, Hulbert, and Trueswell (2008) compared eye movements from Greek and English speakers as they viewed motion events while (*a*) preparing verbal descriptions or (*b*) memorizing the events. During the verbal description task, speakers' eyes rapidly focused on the event components typically encoded in their native language, generating significant cross-language differences even during the first second of motion onset. However, when freely inspecting ongoing events (memorization task), people allocated attention similarly regardless of the language they spoke. Differences between language groups arose only after the motion stopped, such that participants spontaneously studied those aspects of the scene that their language did not routinely encode in verbs (e.g., English speakers were more likely to focus on the path and Greek speakers on the manner of the event). These findings indicate that attention allocation during event perception is not affected by the perceiver's native language; effects of language arise only when linguistic forms are recruited to achieve the task, such as when committing facts to memory. A separate study confirmed that the linguistic intrusions observed at late stages of event inspection in the memory task of Papafragou et al. (2008) disappear under conditions of linguistic interference (e.g., if people are asked to inspect events while repeating back strings of numbers) but persist under conditions of nonlinguistic interference (e.g., if people view events while tapping sounds they hear; Trueswell & Papafragou, 2010). Together, these studies suggest that cross-linguistic differences do not invade (nonlinguistic) event apprehension. Nevertheless, language (if available) can be recruited to help event encoding, particularly in tasks that involve heavy cognitive load.

Spatial Frames of Reference

Certain linguistic communities (e.g., Tenejapan Mayans) customarily use an externally referenced ("absolute") spatial-coordinate system to refer to nearby directions and positions ("to the north"); others (e.g., Dutch speakers) typically use a viewer-perspective ("relative") system ("to the left"). Brown and Levinson (1993) and Pederson et al. (1998) claim that these linguistic practices affect spatial

reasoning in language-specific ways. In one of their experiments, Tenejapan Mayan and Dutch subjects were presented with an array of objects (toy animals) on a tabletop; after a brief delay, subjects were taken to the opposite side of a new table (they were effectively rotated 180 degrees), handed the toys, and asked to reproduce the array “in the same way as before.” The overwhelming majority of Tenejapan (“absolute”) speakers rearranged the objects so that they were heading in the same cardinal direction after rotation, while Dutch (“relative”) speakers massively preferred to rearrange the objects in terms of left-right directionality. This covariation of linguistic terminology and spatial reasoning seems to provide compelling evidence for linguistic influences on nonlinguistic cognition.¹¹

However, as so often in this literature, it is quite hard to disentangle cause and effect. For instance, it is possible that that the Tenejapan and Dutch groups think about space differently because their languages pattern differently, but it is just as possible that the two linguistic-cultural groups developed different spatial-orientational vocabulary to reflect (rather than cause) differences in their spatial reasoning strategies. Li and Gleitman (2002) investigated this second position. They noted that absolute spatial terminology is widely used in many English-speaking communities whose environment is geographically constrained and includes large stable landmarks such as oceans and looming mountains. For instance, the absolute terms *uptown*, *downtown*, and *crosstown* (referring to North, South, and East-West) are widely used to describe and navigate in the space of Manhattan Island; Chicagoans regularly make absolute reference to the lake; and so on. It is quite possible, then, that the presence/absence of stable landmark information, rather than language spoken, influences the choice of absolute versus spatial-coordinate frameworks. After all, the influence of such landmark information on spatial reasoning has been demonstrated with nonlinguistic (rats; Restle, 1957) and prelinguistic (infants; Acredolo & Evans, 1980) creatures.

To examine this possibility, Li and Gleitman replicated Brown and Levinson’s rotation task with English speakers, but they manipulated the presence/absence of landmark cues in the testing area. The result, just as for the rats and the infants, was that English-speaking adults respond absolutely in the presence of landmark information (after rotation, they set up the animals going in the same

cardinal direction) and relatively when it is withheld (in this case, they set up the animals going in the same body-relative direction).

More recent findings suggest that the spatial reasoning findings from these investigators are again language-on-language effects, the result of differing understanding of the instruction to make an array “the same” after rotation. Subjects should interpret this blatantly ambiguous instruction egocentrically if common linguistic usage in the language is of “left” and “right,” as in English, but geocentrically if common linguistic usage is of “east” or “west” as in Tseltal. But what should happen if the situation is not ambiguous, that is, if by the nature of the task it requires either one of these solution types or the other? If the subjects’ capacity to reason spatially has been permanently “transformed” by a lifetime of linguistic habit, there should be some cost—increased errorfulness or slowed responding, for instance—in a task that requires the style of reasoning that mismatches the linguistic encoding. Li, Abarbanell, Gleitman, and Papafragou (2011) experimented with such nonambiguous versions of the spatial rotation tasks, yielding the finding that all cross-linguistic differences disappeared. Tseltal-speaking individuals solved these unambiguous rotation tasks at least as well (often better) when they required egocentric strategies as when they required geocentric strategies.

Flexibility in spatial reasoning when linguistic pragmatics do not enter into the task demands should come as little surprise. The ability to navigate in space is hardwired in the brain of moving creatures, including bees and ants; for all of these organisms, reliable orientation and navigation in space is crucial for survival (Gallistel, 1990); not surprisingly, neurobiological evidence from humans and other species indicates that the brain routinely uses a multiplicity of coordinate frameworks in coding for the position of objects in order to prepare for directed action (Gallistel, 2002). It would be pretty amazing if, among all the creatures that walk, fly, and crawl on the earth, only humans in virtue of acquiring a particular language lose the ability to use both absolute and relative spatial-coordinate frameworks flexibly.

Evidentiality

One of Whorf’s most interesting conjectures concerned the possible effects of evidentials (linguistic markers of information source) on the nature of

thought. Whorf pointed out that Hopi—unlike English—marked evidential distinctions in its complementizer system. Comparing the sentences *I see that it is red* versus *I see that it is new*, he remarked:

We fuse two quite different types of relationship into a vague sort of connection expressed by ‘that,’ whereas the Hopi indicates that in the first case seeing presents a sensation ‘red,’ and in the second that seeing presents unspecified evidence for which is drawn the inference of newness. (Whorf, 1956, p. 85)

Whorf concluded that this grammatical feature was bound to make certain conceptual distinctions easier to draw for the Hopi speaker because of the force of habitual linguistic practices.

Papafragou, Li, Choi, and Han (2007) investigated this proposal. They compared English (which mostly marks evidentiality lexically: “I *saw/heard/inferred* that John left”) to Korean (where evidentiality is encoded through a set of dedicated morphemes). There is evidence that such morphemes are produced early by children learning Korean (Choi, 1995). Papafragou et al. therefore asked whether Korean children develop the relevant conceptual distinctions earlier and with greater reliability than learners of English, which does not grammatically encode this distinction. In a series of experiments, they compared the acquisition of nonlinguistic distinctions between sources of evidence in 3- and 4-year-olds learning English or Korean: no difference in nonlinguistic reasoning in these regards was found between the English and Korean group. For instance, children in both linguistic groups were equally good at reporting how they found out about the contents of a container (e.g., by looking inside or by being told); both groups were also able to attribute knowledge of the contents of a container to a character who had looked inside but not to another character who had had no visual access to its content. Furthermore, Korean learners were more advanced in their nonlinguistic knowledge of sources of information than in their knowledge of the meaning of linguistic evidentials. In this case, then, learned linguistic categories do not seem to serve as “a guide” for the individual’s nonlinguistic categories in the way that Whorf and several later commentators (e.g., Levinson, 2003) have conjectured. Rather, the acquisition of linguistically encoded distinctions seems to follow (and build on) the conceptual understanding of evidential distinctions. The conceptual understanding itself appears to proceed similarly across

diverse language-learning populations. Similar data have recently been obtained from Turkish, where the acquisition of evidential morphology seems to lag behind nonlinguistic knowledge about sources of information (Ozturk & Papafragou, 2008).

Time

So far we have focused on grammatical and lexical properties of linguistic systems and their possible effects on conceptual structure. Here we consider another aspect of languages as expressive systems: their systematically differing use of certain networks of metaphor—specifically, metaphor for talking about time. English speakers predominantly talk about time as if it were horizontal (*one pushes deadlines back, expects good times ahead, or moves meetings forward*). Boroditsky (2001) reports that Mandarin speakers more usually talk about time in terms of a vertical axis (they use the Mandarin equivalents of *up* and *down* to refer to the order of events, weeks, or months). Boroditsky showed that these differences predict aspects of temporal reasoning by speakers of these two languages. In one of her manipulations, subjects were shown two objects in vertical arrangement, say, one fish following another one downward as they heard something like *The black fish is winning*. After this vertically oriented prime, Mandarin speakers were faster to confirm or disconfirm temporal propositions (e.g., *March comes earlier than April*) than if they were shown the fish in a horizontal array. The reverse was true for English speakers. Boroditsky concluded that spatiotemporal metaphors in language affect how people reason about time. She has suggested, more generally, that such systematic linguistic metaphors are important in shaping habitual patterns of thought.

However, these results are again more complex than they seem at first glance. For one thing, and as Boroditsky acknowledges, vertical metaphors of time are by no means absent from ordinary English speech (e.g., *I have a deadline coming up; this recipe came down to me from my grandmother*), though they are said to be more sporadic than in Mandarin. Other laboratories have failed to replicate the original finding (January & Kako, 2007). Moreover, Chen (2007) has disputed the phenomenon altogether, failing to find predominance of the vertical metaphor in a corpus analysis of Taiwanese newspapers. Assuming, though, that the difference does hold up in everyday speech contexts, it is a subtle cross-linguistic difference of degree, rather than a principled opposition.

In fact, the most telling finding reported in these studies is that the apparent inculcation of a generalization over a lifetime is easily erased—in fact, actually reversed—in a matter of minutes: Boroditsky explained to her English-speaking subjects how to talk about time vertically, as in Mandarin, and gave them several practice trials. After this training, the English speakers exhibited the vertical (rather than the former horizontal) priming effect. Apparently, 15 minutes of training on the vertical overcame and completely reversed 20+ years of the habitual use of the horizontal in these speakers. The effects of metaphor, it seems, are transient and fluid, without long-term influence on the nature of conceptualization or its implicit deployment to evaluate propositions in real time. Again, these results are as predicted under a processing—language-on-language—account, in which there are immediate effects on memory (here, repetition and recency effects), but no permanent reorganization of “thought.”

Number

Prelinguistic infants and nonhuman primates share an ability to represent both exact numerosities for very small sets (roughly up to three objects) and approximate numerosities for larger sets (Dehaene, 1997). Human adults possess a third system for representing number, which allows for the representation of exact numerosities for large sets, has (in principle) no upper bound on set size, and can support the comparison of numerosities of different sets as well as processes of addition and subtraction. Crucially, this system is *generative*, since it possesses a rule for creating successive integers (the successor function) and is thus characterized by discrete infinity.

How do young children become capable of using this uniquely human number system? One powerful answer is that the basic principles underlying the adult number system are innate; gaining access to these principles thus gives children a way of grasping the infinitely discrete nature of natural numbers, as manifested by their ability to use verbal counting (Gelman & Gallistel, 1978; also see Opfer & Siegler, Chapter 30). Other researchers propose that children come to acquire the adult number system by conjoining properties of the two prelinguistic number systems via natural language. Specifically, they propose that grasping the *linguistic* properties of number words (e.g., their role in verbal counting, or their semantic relations to quantifiers such as *few*, *all*, *many*, *most*; see Spelke & Tsivkin, 2001a and Bloom, 1994b; Carey, 2001, respectively) enables children

to put together elements of the two previously available number systems in order to create a new, generative number faculty. In Bloom's (1994b, p. 186) words, “in the course of development, children ‘bootstrap’ a generative understanding of number out of the productive syntactic and morphological structures available in the counting system.”

For instance, upon hearing the number words in a counting context, children realize that these words map onto both specific representations delivered by the exact-numerosities calculator and inexact representations delivered by the approximator device. By conjoining properties of these two systems, children gain insight into the properties of the adult conception of number (e.g., that each of the number words picks out an exact set of entities, that adding or subtracting exactly one object changes number, etc.). Ultimately, it is hypothesized that this process enables the child to compute exact numerosities even for large sets (such as *seven* or *twenty-three*)—an ability which was not afforded by either one of the prelinguistic calculation systems.

Spelke and Tsivkin (2001a, b) experimentally investigated the thesis that language contributes to exact large-number calculations. In their studies, bilinguals who were trained on arithmetic problems in a single language and later tested on them were faster on large-number arithmetic if tested in the training language; however, no such advantage of the training language appeared with estimation problems. The conclusion from this and related experiments was that the particular natural language is the vehicle of thought concerning large exact numbers but not about approximate numerosities. Such findings, as Spelke and her collaborators have emphasized, can be part of the explanation of the special “smartness” of humans (see also Penn & Povinelli, Chapter 27 for similar views). Higher animals, like humans, can reason to some degree about approximate numerosity, but not about exact numbers. Beyond this shared core knowledge, however, humans have language. If language is a required causal factor in exact number knowledge, this in principle could explain the gulf between creatures like us and creatures like them. In support of the dependence of the exact number system on natural language, recent findings have shown that members of the Pirahá community that lack number words and a counting system seem unable to compute exact large numerosities (Gordon, 2004).

How plausible is the view that the adult number faculty presupposes linguistic mediation? Recall

that, on this view, children infer the generative structure of number from the generative structure of grammar when they hear others counting. However, counting systems vary cross-linguistically, and in a language like English, their recursive properties are not really obvious from the outset. Specifically, until number eleven, the English counting system presents no evidence of regularity, much less of generativity: A child hearing *one, two, three, four, five, six*, and up to *eleven* would have no reason to assume—based on properties of form—that the corresponding numbers are lawfully related (namely, that they successively increase by one). For larger numbers, the system is more regular, even though not fully recursive due to the presence of several idiosyncratic features (e.g., one can say *eighteen* or *nineteen* but not *teneen* for twenty). In sum, it is not so clear how the “productive syntactic and morphological structures available in the counting system” will provide systematic examples of discrete infinity that can then be imported into number cognition.

Can properties of other natural language expressions bootstrap a generative understanding? Quantifiers have been proposed as a possible candidate (Carey, 2001). However, familiar quantifiers lack the hallmark properties of the number system: They are not strictly ordered with respect to one another and their generation is not governed by the successor function. In fact, several quantifiers presuppose the computation of cardinality of sets: for example, *neither* and *both* apply only to sets of two items (Barwise & Cooper, 1981; Keenan & Stavi, 1986). Moreover, quantifiers and numbers compose in quite different ways. For example, the expression *most men and women* cannot be interpreted to mean a large majority of the men and much less than half the women. In light of the semantic disparities between the quantifier and the integer systems, it is hard to see how one could bootstrap the semantics of the one from the other.

Experimental findings suggest, moreover, that young children understand certain semantic properties of number words well before they know those of quantifiers. One case involves the scalar interpretation of these terms. In one experiment, Papafragou and Musolino (2003) had 5-year-old children watch as three horses are shown jumping over a fence. The children would not accept *Two of the horses jumped over the fence* as an adequate description of that event (even though it is necessarily true that if three horses jumped, then certainly two did). But at the same age, they would accept *Some of the horses jumped over the*

fence as an adequate description, even though it is again true that all of the horses jumped. In another experiment, Hurewitz, Papafragou, Gleitman, and Gelman (2006) found that 3-year-olds understand certain semantic properties of number words such as *two* and *four* well before they know those of quantifiers such as *some* and *all*. It seems, then, that the linguistic systems of number and natural language quantification are developing rather independently. If anything, the children seem more advanced in knowledge of the meaning of number words than quantifiers, so it is hard to see how the semantics of the former lexical type is to be bootstrapped from the semantics of the latter.

How then are we to interpret the fact that linguistic number words seem to be crucially implicated in nonlinguistic number cognition (Gordon, 2004; Spelke & Tsivkin, 2001a, b)? One promising approach is to consider number words as a method for online encoding, storage, and manipulation of numerical information that complements, rather than altering or replacing, nonverbal representations. Evidence for this claim comes from recent studies that retested the Pirahá population in tasks used by Gordon (Frank, Everett, Fedorenko, & Gibson, 2008). Pirahá speakers were able to perform exact matches with large numbers of objects perfectly, but, as previously reported, they were inaccurate on matching tasks involving memory. Other studies showed that English-speaking participants behave similarly to the Pirahá population on large number tasks when verbal number representations are unavailable due to verbal interference (Frank, Fedorenko, & Gibson, 2008). Nicaraguan signers who have incomplete or nonexistent knowledge of the recursive count list show a similar pattern of impairments (Flaherty & Senghas, 2007). Together, these data are consistent with the hypothesis that verbal mechanisms are necessary for learning and remembering large exact quantities—an online mnemonic effect of language of a sort we have already discussed.

Orientation

A final domain that we will discuss is spatial orientation. Cheng and Gallistel (1984) found that rats rely on geometric information to reorient themselves in a rectangular space, and they seem incapable of integrating geometrical with nongeometrical properties (e.g., color, smell, etc.) in searching for a hidden object. If they see food hidden at the corner of a long and a short wall, they will search equally

at either of the two such walls of a rectangular space after disorientation; this is so even if these corners are distinguishable by one of the long walls being painted blue or having a special smell. Hermer and Spelke (1994, 1996) reported a very similar difficulty in young children. Both animals and young children can navigate and reorient by the use of either geometric or nongeometric cues; it is integrating across the cue types that makes the trouble. These difficulties are overcome by older children and adults who are able, for instance, to go straight to the corner formed by a long wall to the left and a short blue wall to the right. Hermer and Spelke found that success in these tasks was significantly predicted by the spontaneous combination of spatial vocabulary and object properties such as color within a single phrase (e.g., *to the left of the blue wall*).¹² Later experiments (Hermer-Vasquez, Spelke, & Katsnelson, 1999) revealed that adults who were asked to shadow speech had more difficulty in these orientation tasks than adults who were asked to shadow a rhythm with their hands; however, verbal shadowing did not disrupt subjects' performance in tasks which required the use of nongeometric information only. The conclusion was that speech-shadowing, unlike rhythm-shadowing, by taking up linguistic resources, blocked the integration of geometrical and object properties that is required to solve complex orientation tasks. In short, success at the task seems to require encoding of the relevant terms in a specifically linguistic format.

In an influential review article, Carruthers (2002) suggests even more strongly that in number, space, and perhaps other domains, language is the medium of intermodular communication, a format in which representations from different domains can be combined in order to create novel concepts. However, on standard assumptions about modularity, modules are characterized as computational systems with their own proprietary vocabulary and combinatorial rules. Since language itself is a module in this sense, its computations and properties (e.g., generativity, compositionality) cannot be "transferred" to other modules, because they are defined over—and can only apply to—language-internal representations. One way out of this conundrum is to give up the assumption that language is—on the appropriate level—modular:

Language may serve as a medium for this conjunction...because it is a domain-general, combinatorial system to which the representations delivered by the child's...[domain-specific]

nonverbal systems can be mapped. (Spelke & Tsivkin, 2001b, p. 84)

And:

Language is constitutively involved in (some kinds of) human thinking. Specifically, language is the vehicle of non-modular, non-domain-specific, conceptual thinking which integrates the results of modular thinking. (Carruthers, 2002, p. 666)

On this view, the output of the linguistic system just IS Mentalese: There is no other level of representation in which the information *to the left of the blue wall* can be entertained. This picture of language is novel in many respects. In the first place, replacing Mentalese with a linguistic representation challenges existing theories of language production and comprehension. Traditionally, the production of sentences is assumed to begin by entertaining the corresponding thought, which then mobilizes the appropriate linguistic resources for its expression (e.g., Levelt, 1989). On some proposals, however,

We cannot accept that the production of a sentence 'The toy is to the left of the blue wall' begins with a tokening of the thought THE TOY IS TO THE LEFT OF THE BLUE WALL (in Mentalese), since our hypothesis is that such a thought cannot be entertained independently of being framed in a natural language. (Carruthers, 2002, p. 668)

Inversely, language comprehension is classically taken to unpack linguistic representations into mental representations, which can then trigger further inferences. But in Carruthers' proposal, after hearing *The toy is to the left of the blue wall*, the interpretive device cannot decode the message into the corresponding thought, since there is no level of Mentalese independent of language in which the constituents are lawfully connected to each other. Interpretation can only dismantle the utterance and send its concepts back to the geometric and landmark modules to be processed. In this sense, understanding an utterance such as *The picture is to the right of the red wall* turns out to be a very different process than understanding superficially similar utterances such as *The picture is to the right of the wall* or *The picture is on the red wall* (which do not, on this account, require cross-domain integration).

Furthermore, if language is to serve as a domain for cross-module integration, then the lexical resources of each language become crucial for conceptual combination. For instance, lexical gaps in

the language will block conceptual integration, since there would be no relevant words to be inserted into the linguistic string. As we have discussed at length, color terms vary across languages (Kay & Regier, 2002); more relevantly, not all languages have terms for *left* and *right* (Levinson, 1996). It follows that speakers of these languages should fail to combine geometric and object properties in the same way as do English speakers in order to recover from disorientation. In other words, depending on the spatial vocabulary available in their language, disoriented adults may behave either like Spelke and Tsivkin's English-speaking population or like prelinguistic infants and rats. This prediction, although merely carrying the original proposal to its apparent logical conclusion, is quite radical: It allows a striking discontinuity among members of the human species, contingent not on the presence or absence of human language and its combinatorial powers (as the original experiments seem to suggest), or even on cultural and educational differences, but on vagaries of the lexicon in individual linguistic systems.

Despite its radical entailments, there is a sense in which Spelke's proposal to interpret concept configurations on the basis of the combinatorics of natural language can be construed as decidedly nativist. In fact, we so construe it. Spelke's proposal requires that humans be equipped with the ability to construct novel structured syntactic representations, insert lexical concepts at the terminal nodes of such representations (*left*, *blue*, etc.), and interpret the outcome on the basis of familiar rules of semantic composition (*to the left of the blue wall*). In other words, humans are granted principled knowledge of how phrasal meaning is determined by lexical units and the way they are composed into structured configurations. That is, what is granted is the ability to read the semantics off of phrase structure trees. Furthermore, the assumption is that this knowledge is not itself attained through learning but belongs to the in-built properties of the human language device.

But notice that granting humans the core ability to build and interpret phrase structures is granting them quite a lot. Exactly these presuppositions have been the hallmark of the nativist program in linguistics and language acquisition (Chomsky, 1957; Gleitman, 1990; Jackendoff, 1990; Lidz et al., 2003; Pinker, 1984) and the target of vigorous dissent elsewhere (Goldberg, 1995; Tomasello, 2000). To the extent that Spelke and Tsivkin's arguments

about language and cognition rely on the combinatorial and generative powers of language, they make deep commitments to abstract (and unlearnable) syntactic principles and their semantic reflexes. Notice in this regard that since these authors hold that *any* natural language will do as the source and vehicle for the required inferences, the principles at work here must be abstract enough to wash out the diverse surface-structural realizations of *to the left of the blue wall* in the languages of the world. An organism with such principles in place could—*independently of particular experiences*—generate and *systematically* comprehend novel linguistic strings with meanings predictable from the internal organization of those strings—and, for different but related reasons, *just as systematically* fail to understand other strings such as *to the left of the blue idea*. We would be among the very last to deny such a proposal in its general form. We agree that there are universal aspects of the syntax-semantics interface. Whether these derive from or augment the combinatorial powers of thought is the question at issue here.

Recent developmental studies from Dessalegn and Landau (2008) offer useful ways to understand the issue just raised (see also Landau et al. 2009). These investigators studied 4-year-olds' ability to keep track of two features of a visual array simultaneously: color and position. Classic work from Treisman and Schmidt (1982) has shown that such visual features are initially processed independently, so that under rapid presentation, a red "O" next to a green "L" might be reported as a green O even by adults. Young children are even more prone to such errors, often giving mirror-image responses to, for example, a square green on its left side and red on its right. Directions such as "Look very hard" or "Look! The red is touching the green" do not reduce the prevalence of such errors. But subjects told "Look! The red is on the right" improve dramatically. Landau and colleagues point out that this finding in itself isn't very surprising—except that they show that these preschoolers did not have a stable grasp of the meanings of the terms *left* versus *right*, when tested for this separately. Yet their partial, possibly quite vague, sensitivity to these egocentric spatial terms was enough to influence perceptual performance "in the moment." Two properties of these findings further support the interpretation that applies to most of the results we have reported. First, the linguistic influence is highly transient—a matter of milliseconds. Second, the

effect, presumably like those of Hermer and Spelke, is independent of *which* language is being tested. Rather, as Landau and colleagues put it, there is a momentary “enhancement” of cognitive processing in the presence of very specific linguistic labeling.

Conclusions and Future Directions

We have just reviewed several topics within the burgeoning psychological and anthropological literature that are seen as revealing causal effects of language on thought, in senses indebted to Sapir and Whorf. We began discussion with the many difficulties involved in radical versions of the linguistic “determinism” position, including the fact that language seems to underspecify thought, and to diverge from it as to the treatment of ambiguity, paraphrase, and deictic reference. Moreover, there is ample evidence that several forms of cognitive organization are independent of language: Infants who have no language are able to entertain relatively complex thoughts; for that matter, they can learn languages or even invent them when the need arises (Feldman, Goldin-Meadow, & Gleitman, 1978; Goldin-Meadow, 2003; Senghas, Coppola, Newport, & Suppala, 1997); many bilinguals as a matter of course “code-switch” between their known languages even within a single sentence (Joshi, 1985); aphasics sometimes exhibit impressive propositional thinking (Varley & Siegal, 2000); animals can form representations of space, artifacts, and perhaps even mental states without linguistic crutches (Gallistel, 1990; Hare, Call, & Tomasello, 2001). All these nonlinguistic instances of thinking and reasoning (also see Hegarty & Stull, Chapter 31) dispose of the extravagant idea that language just “is” thought.

However, throughout this chapter we have surveyed approximately half a century of investigation in many cognitive-perceptual domains that document systematic population differences in behavior, attributable to the particular language spoken. Consistent and widespread as these findings have been, there is little scientific consensus on their interpretation. Quite the contrary, recent positions range from those holding that specific words or language structures cause “radical restructuring of cognition” (e.g., Majid, Bowerman, Kita, Haun, & Levinson, 2004) to those that maintain—based on much the same kinds of findings—that there is a “remarkable independence of language and thought” (e.g., Heider & Oliver, 1972; Jameson & Hurvich, 1978). To approach these issues, it is instructive to

reconsider the following three steps that have always characterized the relevant research program:

- (1) *Identify a difference* between two languages, in sound, word, or structure.
- (2) *Demonstrate a concordant cognitive or perceptual difference* between speakers of the languages identified in (1).
- (3) *Conclude that, at least in some cases, (1) caused (2)* rather than the other way round.

Though there is sometimes interpretive difficulty at step (3)—recall Eskimos in the snow—the major problem is to disambiguate the source of the differences discovered at step (2). To do so, investigators either compare results when a linguistic response is or is not part of the task (e.g., Jameson & Hurvich, 1978; Li et al., 2009; Papafragou et al., 2008, or that do or do not interfere with simultaneous linguistic functioning (e.g., Frank et al., 2008; Kay & Kempton, 1984; Trueswell & Papafragou, 2010; Winauer et al., 2007); or where hemispheric effects, implicating or not implicating language areas in the brain, can be selectively measured (e.g., Regier et al., 2010). The cross-language differences are usually diminished or disappear under those conditions where language is selectively excluded. Traditionally, investigators have concluded from this pattern of results that language categories do not penetrate deeply into nonlinguistic thought, and therefore that the Sapir-Whorf-conjecture has been deflated or discredited altogether.

But surprisingly, recent commentary has sometimes stood this logic on its head. Interpretation of these same patterns has been to the effect that, when behavioral differences arise if and only if language *is* implicated in the task, this is evidence *supporting* the Sapir-Whorf thesis, that is, vindicating the view that language causally impacts and transforms thought. Here is L. Boroditsky (2010) in a recent commentary on the color-category literature:

...disrupting people’s ability to use language while they are making colour judgments eliminates the cross-linguistic differences. This demonstrates that language *per se* plays a causal role, meddling in basic perceptual decisions as they happen.

Thus, at first glance, investigators are in the quandary of fact-immune theorizing, in which no matter how the results of experimentation turn out, the hypothesis is confirmed. As Regier et al. (2010) put this in a recent review, such findings

... act as a sort of Rorschach test. Those who "want" the Whorf hypothesis to be true can point to the fact that the manipulation clearly implicates language. At the same time, those who "want" the hypothesis to be false can point to how easy it is to eliminate effects of language on perception, and argue on that basis that Whorfian effects are superficial and transient. (p. 179)

In the present chapter, we have understood the literature in a third way, one that situates the findings in each of the domains reviewed squarely within the "ordinary" psycholinguistic literature, as "language-on-language" effects: language-specific patterns of cognitive performance are a product of the online language processing that occurs during problem solving. These patterns are indeed transient in the sense that they do not change the nature of the domain itself (pace Whorf, 1956, and Pederson et al., 1998), but are by no means superficial. In some cases, such effects are outcomes of linguistic information handling, as these emerge online, in the course of understanding the verbal instructions in a cognitive task. For instance, because of the differential frequencies, and so on, of linguistic categories across languages, slightly different problems may be posed to the processing apparatus of speakers of different languages by what appear to be "identical" verbal instructions in an experiment (see discussion of Imai & Gentner's, 1997, results on object individuation). In other cases, linguistic information may be used online to recode nonlinguistic stimuli even if the task requires no use of language. This is particularly likely to happen in tasks with high cognitive load (Trueswell & Papafragou, 2010), because language is an efficient way to represent and store information. In neither case of linguistic intrusion does language reshape or replace other cognitive formats of representation, but it does offer a mode of information processing that is often preferentially invoked during cognitive activity (for related statements, see Fisher & Gleitman, 2002; Papafragou et al., 2002; Papafragou et al., 2008; Trueswell & Papafragou, 2010).

Other well-known findings about the role of language in cognition are consistent with this view. For example, a major series of developmental studies demonstrate that a new linguistic label "invites" the learner to attend to certain types of classification criteria over others or to promote them in prominence. Markman and Hutchinson (1984) found that if one shows a 2-year-old a new object and says, *See this one; find another one*, the child typically reaches for

something that has a spatial or encyclopedic relation to the original object (e.g., finding a bone to go with the dog). But if one uses a new word (*See this fendle, find another fendle*), the child typically looks for something from the same category (e.g., finding another dog to go with the first dog). Balaban and Waxman (1997) showed that labeling can facilitate categorization in infants as young as 9 months (cf. Xu, 2002). Beyond categorization, labeling has been shown to guide infants' inductive inference (e.g., expectations about nonobvious properties of novel objects), even more so than perceptual similarity (Welder & Graham, 2001). Other recent experimentation shows that labeling may help children solve spatial tasks by pointing to specific systems of spatial relations (Loewenstein & Gentner, 2005). For learners, then, the presence of linguistic labels constrains criteria for categorization and serves to foreground a *codable* category out of all the possible categories a stimulus could be said to belong to. Here, as well, the presence of linguistic labels does not intervene in the sense of replacing or reshaping underlying (nonlinguistic) categories; rather, it offers an alternative, efficient system of encoding, organizing, and remembering experience.

Acknowledgments

Preparation of this chapter has been supported in part by grant 5-R01-HD55498 from the National Institutes of Health to A.P. and John Trueswell and in part by grant 1-R01-HD37507 from the National Institutes of Health to L.R.G and John Trueswell.

Notes

1. "Gilead then cut Ephraim off from the fords of the Jordan, and whenever Ephraimite fugitives said, 'Let me cross,' the men of Gilead would ask, 'Are you an Ephraimite?' If he said, 'No,' they then said, 'Very well, say "Shibboleth."' If anyone said, "Sibboleth," because he could not pronounce it, then they would seize him and kill him by the fords of the Jordan. Forty-two thousand Ephraimites fell on this occasion" (Judges 12-5-6; as cited in Wikipedia).

2. Whorf's own position, and specific claims, on all the matters discussed in this chapter was often metaphorical, highly nuanced, and to some extent inconsistent across his body of work. Sometimes his concentration was on cultural differences as reflected in language rather than on language as tailor of culture. Sometimes he asserted, but sometimes rejected, the idea that particular words, word classes, or grammatical devices ("surface" facts about language) were his intended causal vehicles of mental categories and functions. Owing to this partial inconsistency, perhaps a common property of scientific views in their earliest formulations, an industry of interpreting Whorf —both by his defenders and detractors—has grown up, and it is often heated. Our aim is to explicate the theoretical positions ("Whorfianism") that are indebted in one or another way to this thinker, not to present an intellectual biography of Whorf himself.

3. We thank Jerry Fodor for discussion of these issues.

4. In one experimental demonstration, subjects were asked: *When an airplane crashes, where should the survivors be buried?* They rarely noticed the meaning discrepancy in the question (Barton & Sanford, 1993).

5. The similarity test may not be decisive for this case, as Malt, Sloman, and Gennari (2003) as well as Smith et al. (2001), among others, have pointed out. Similarity judgments as the measuring instrument could be systematically masking various nonperceptual determinants of organization in a semantic-conceptual domain, some of these potentially language-caused. Over the course of this essay, we will return to consider other domains and other psychological measures. For further discussion of the sometimes arbitrary and linguistically varying nature of the lexicon, even in languages which are typologically and historically closely related, see Kay (1996). He points out, for example, that English speakers use *screwdriver* while the Germans use *Schraubenzieher* (literally, “screwpuller”), and the French *tournevis* (literally, “screwturner”) for the same purposes; our turnpike exit-entry points are marked *exit*, whereas the Brazilians have *entradas*; and so forth.

6. These results are fairly recent, and a number of follow-up studies suggest that the picture that finally emerges may be more complicated than foreseen in Gilbert et al. For instance, Lindsey et al. (2010) report that some desaturated highly codable colors (notably, certain pinks) are not rapidly identified. Liu et al. (2009) do replicate the between-category advantage finding of Gilbert et al., but, critically, not the hemispheric advantage. If so, the suggestion is that labeling practice is penetrating to the level of nonlinguistic cognition. Roberson and colleagues adopt this very view (e.g., Roberson, 2005; Roberson, Davies, & Davidoff, 2000; Roberson, Davidoff, Davies, & Shapiro, 2005), reporting, for example, that Berinmo speakers (members of another relatively isolated Papua New Guinea tribe) were better at recognizing and remembering best examples of their own linguistic categories than color labels than the best examples of English color labels. They use such results to claim that color naming is entirely arbitrary from the point of view of perception, being solely a matter of linguistic labeling practices (for a response, see Kay & Regier, 2007).

7. This argument is not easy. After all, one might argue that English is a classifier language much like Yucatec Mayan or Japanese, that is, that all its words start out as mass nouns and become countable entities only through adding the classifiers *the* and *a* (compare *brick* the substance to *a brick*, the object). However, detailed linguistic analysis suggests that there is a genuine typological difference here; see Slobin, 2001; Chierchia, 1998; Krifka, 1995; Lucy & Gaskins, 2001, for discussion). The question is whether, since all of the languages formally mark the mass/count distinction in one way or another, the difference in particular linguistic means could plausibly rebound to impact ontology.

8. We should point out that this hint is itself at best a weak one, another reason why the observed interpretive difference for Japanese and English speakers, even at the perceptual midline, is also weak. Notoriously, English often violates the semantic generalization linking mass noun morphology with substancehood (compare, e.g., *footwear*; *silverware*; *furniture*).

9. Subsequent analysis of the linguistic data revealed that Greek speakers were more likely to include manner of motion in their verbal descriptions when manner was unexpected or non-inferable, while English speakers included manner information regardless of inferability (Papafragou et al. 2006). This suggests that speakers may monitor harder-to-encode event components and choose to include them in their utterances when especially

informative. This finding reinforces the conclusion that verbally encoded aspects of events vastly underdetermine the subtleties of event cognition. As Brown and Dell had shown earlier (1987), English actually shows the same tendency, but more probabilistically: English speakers are less likely to express an inferable instrument, for example, to say, “He stabbed him with a knife” than a noninferable one (“He stabbed him with an ice cutter”).

10. In another demonstration of this language-on-language effect, Naigles and Terrazas (1998) asked subjects to describe and categorize videotaped scenes, for example, of a girl skipping toward a tree. They found that Spanish- and English-speaking adults differed in their preferred interpretations of new (nonsense) motion verbs in manner-biasing (*She's kradding toward the tree* or *Ella está mecanando hacia el árbol*) or path-biasing (*She's kradding the tree* or *Ella está mecanando el árbol*) sentence structures. The interpretations were heavily influenced by syntactic structure. But judgments also reflected the preponderance of verbs in each language—Spanish speakers gave more path interpretations, and English speakers gave more manner interpretations. Similar effects of language-specific lexical practices on presumed verb extension have been found for children (Papafragou & Selimis, 2010a).

11. It might seem perverse to hold (as Levinson and colleagues do) that it is “lacking ‘left,’” rather than “having ‘east,’” that explains the navigational skills of the Mayans, and the relative lack of such skills in speakers of most European languages. The reason, presumably, is that all languages have and widely use vocabulary for geocentric location and direction, so to point to one language’s geocentric vocabulary would not account for the presumptive behavioral difference in navigational skill. Therefore, by hypothesis, it must be the mere presence of the alternate vocabulary (of body-centered terms) that’s doing the damage. Here L. Boroditsky (2010) makes this position explicit: “For example, unlike English, many languages do not use words like ‘left’ and ‘right’ and instead put everything in terms of cardinal directions, requiring their speakers to say things like ‘there’s an ant on your south-west leg.’ As a result, speakers of such languages are remarkably good at staying oriented (even in unfamiliar places or inside buildings) and perform feats of navigation that seem superhuman to English speakers. In this case, just a few words in a language make a big difference in what cognitive abilities their speakers develop.”

12. Further studies show that success in this task among young children is sensitive to the size of the room—in a large room, more 4-year-olds succeed in combining geometric and landmark information (Learmonth, Nadel, & Newcombe, 2002). Also, when adults are warned about the parameters of the task, they are able to fall back on alternative representational strategies (Ratliff & Newcombe, 2008). Moreover, it is claimed that other species (chickens, monkeys) can use both types of information when disoriented (Gouteux, Thinus-Blanc, & Vauclair, 2001; Vallortigara, Zanforlin, & Pasti, 1990).

References

- Acredolo, L., & Evans, D. (1980). Developmental changes in the effects of landmarks on infant spatial behavior. *Developmental Psychology, 16*, 312–318.
Aslin, R. N. (1981). Experiential influences and sensitive periods in perceptual development: A unified model. In R. N. Aslin, J. R. Alberts, & M. R. Petersen (Eds.), *Development of perception: Psychobiological perspectives, Vol II* (pp. 45–93). New York: Academic Press.
Aslin, R. N., & Pisoni, D. B. (1980). Some developmental processes in speech perception. In G. H. Yeni-Komshian,

- J. F. Kavanagh, & C. A. Ferguson (Eds.), *Child Phonology: Volume 2, Perception* (pp. 67–96). New York: Academic Press.
- Baillargeon, R. (1993). The object concept revisited: New directions in the investigation of infants' physical knowledge. In C. E. Granrud (Ed.), *Carnegie Mellon Symposia on Cognition, Vol. 23: Visual perception and cognition in infancy* (pp. 265–315). Hillsdale, NJ: Erlbaum.
- Baker, M. (2001). *The atoms of language*. New York: Basic Books.
- Balaban, M. T., & Waxman, S. R. (1997). Do words facilitate object categorization in 9-month-old infants? *Journal of Experimental Child Psychology*, 64, 3–26.
- Barton, S. B. & Sanford, A. J. (1993). A case study of anomaly detection: Shallow semantic processing and cohesion establishment. *Memory and Cognition*, 21, 477–487.
- Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4, 159–219.
- Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley: University of California Press.
- Berman, R., & Slobin, D. (Eds.). (1994). Relating events in narrative: A cross-linguistic developmental study. Hillsdale, NJ: Erlbaum.
- Best, C., McRoberts, G., & Sithole, N. (1988). The phonological basis of perceptual loss for nonnative contrasts: Maintenance of discrimination among Zulu clicks by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 345–360.
- Bloch, B. & Trager, G. L. (1942) *Outline of linguistic analysis*. Baltimore, MD: Waverly Press.
- Bloom, P. (1994a). Possible names: The role of syntax-semantics mappings in the acquisition of nominals. *Lingua*, 92, 297–329.
- Bloom, P. (1994b). Generativity within language and other cognitive domains. *Cognition*, 51, 177–189.
- Bloom, P. (2000). *How children learn the meaning of words*. Cambridge, MA: MIT Press.
- Boroditsky, L. (2001). Does language shape thought? Mandarin and English speakers' conception of time. *Cognitive Psychology*, 43, 1–22.
- Boroditsky, L. (2010, December 13). "Pro", Economist on-line debate on language and thought. <http://www.economist.com/debate/days/view/626>
- Bowberman, M. (1996). The origins of children's spatial semantic categories: Cognitive versus linguistic determinants. In J. Gumperz & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 145–176). Cambridge, England: Cambridge University Press.
- Bowberman, M., & Levinson, S. C. (Eds.). (2001a). *Language acquisition and conceptual development*. Cambridge, England: Cambridge University Press.
- Bowberman, M., & Levinson, S. C. (2001b). Introduction. In M. Bowberman & S. C. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 1–16). Cambridge, England: Cambridge University Press.
- Brown, P., & Dell, G. S. (1987). Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology*, 19, 441–472.
- Brown, P., & Levinson, S. C. (1993). "Uphill" and "downhill" in Tzeltal. *Journal of Linguistic Anthropology*, 3, 46–74.
- Brown, R. (1957). Linguistic determinism and the parts of speech. *Journal of Abnormal and Social Psychology*, 55, 1–5.
- Brown, R., & Lenneberg, E. (1954). A study of language and cognition. *Journal of Abnormal and Social Psychology*, 49, 454–462.
- Carey, S. (1982). The child as word learner. In M. Halle, J. Bresnan, & G. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 264–293). Cambridge, MA: MIT Press.
- Carey, S. (2001). Whorf versus continuity theorists: Bringing data to bear on the debate. In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 185–214). Cambridge, England: Cambridge University Press.
- Carey, S. (2008). Math schemata and the origins of number representations. *Behavioral and Brain Sciences*, 31(6), 645–646.
- Carruthers, P. (2002). The cognitive functions of language. *Behavioral and Brain Sciences*, 25, 657–674.
- Chen, J. Y. (2007). Do Chinese and English speakers think about time differently? Failure of replicating Boroditsky (2001). *Cognition*, 104(2), 427–436.
- Cheng, K., & Gallistel, C. R. (1984). Testing the geometric power of an animal's spatial representation. In H. Roitblat, T.G. Bever, & H. Terrace (Eds.), *Animal cognition* (pp. 409–423). Hillsdale, NJ: Erlbaum.
- Chierchia, G. (1998). Reference to kinds across languages. *Natural Language Semantics*, 6, 339–405.
- Choi, S., & Bowerman, M. (1991). Learning to express motion events in English and Korean: The influence of language-specific lexicalization patterns. *Cognition*, 41, 83–121.
- Choi, S. (1995). The development of epistemic sentence-ending modal forms and functions in Korean children. In J. Bybee & S. Fleischman (Eds.), *Modality in grammar and discourse* (pp. 165–204). Amsterdam, Netherlands: John Benjamins.
- Chomsky, N. (1957). *Syntactic structures*. The Hague, Netherlands: Mouton.
- Chomsky, N. (1964) *Current issues in linguistic theory*. The Hague, Netherlands: Mouton.
- Chomsky, N. (1975). *Reflections on language*. New York: Pantheon.
- Clark, H. (1992). *Arenas of language use*. Chicago, IL: University of Chicago Press.
- Dehaene, S. (1997). *The number sense*. New York: Oxford University Press.
- Descartes, R. (1662). *Trait de l'homme*. (E. S. Haldane & G. R. T. Ross, Trans.). Cambridge, England: Cambridge University Press.
- Dessalegn B., & Landau B. (2008). More than meets the eye: The role of language in binding visual properties. *Psychological Science*, 19, 189–195.
- Eimas, P., Siqueland, E., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, 171, 303–306.
- Feldman, H., Goldin-Meadow, S., & Gleitman, L. R. (1978). Beyond Herodotus: The creation of language by linguistically deprived deaf children. In A. Lock (Ed.), *Action, gesture, and symbol: The emergence of language* (pp. 351–414). London: Academic Press.
- Fisher, C. (1996). Structural limits on verb mapping: The role of analogy in children's interpretations of sentences. *Cognitive Psychology*, 31, 41–81.
- Fisher, C., & Gleitman, L. R. (2002). Breaking the linguistic code: Current issues in early language learning. In H. F. Pashler (Series Ed.) & R. Gallistel (Vol. Ed.), *Steven's handbook of experimental psychology, Vol. 1: Learning and motivation* (3rd ed., pp. 445–496). New York: Wiley.
- Flaherty, M., & Senghas, A. (2007). Numerosity and number signs in deaf Nicaraguan adults. In *Proceedings of the 31st Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.
- Fodor, J. (1975). *The language of thought*. New York: Crowell.

- Frank, M. C., Everett, D. L., Fedorenko, E., & Gibson, E., (2008). Number as a cognitive technology: Evidence from Pirahá language and cognition. *Cognition*, 108, 819–824.
- Frank, M. C., Fedorenko, E., & Gibson, E. (2008). Language as a cognitive technology: English-speakers match like Pirahá when you don't let them count. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Gallistel, C. R. (1990). *The organization of learning*. Cambridge, MA: MIT Press.
- Gallistel, C. R. (2002). Language and spatial frames of reference in mind and brain. *Trends in Cognitive Science*, 6, 321–322.
- Gelman, R., & Gallistel, C. R. (1978). *The child's understanding of number*. Cambridge, MA: Harvard University Press.
- Gelman, R., & Spelke, E. (1981). The development of thoughts about animate and inanimate objects: Implications for research on social cognition. In J. H. Flavell & L. Ross (Eds.), *Social cognitive development: Frontiers and possible futures* (pp. 43–66). Cambridge, England: Cambridge University Press.
- Gennari, S., Sloman, S., Malt, B., & Fitch, W. (2002). Motion events in language and cognition. *Cognition*, 83, 49–79.
- Gentner, D., & Boroditsky, L. (2001). Individuation, relativity and early word learning. In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 215–256). Cambridge, England: Cambridge University Press.
- Gentner, D., & Goldin-Meadow, S. (Eds.). (2003). *Language in mind: Advances in the study of language and thought*. Cambridge, MA: MIT Press.
- Gibson, E. J., & Spelke, E. S. (1983). The development of perception. In P. Mussen (Series Ed.) & J. H. Flavell & E. Markman (Eds.), *Handbook of child psychology* (Vol. 3). New York: Wiley.
- Gilbert, A., Regier, T., Kay, P., & Ivry, R. (2006). Whorf hypothesis is supported in the right visual field but not the left. *PNAS*, 103, 489–494.
- Gleitman, L. (1990). The structural sources of verb meaning. *Language Acquisition*, 1, 1–55.
- Gleitman, L. R., January, D., Nappa, R., & Trueswell, J. T. (2007). On the give and take between event apprehension and sentence formulation. *Journal of Memory and Language*, 57(4), 544–569.
- Gleitman, L. R. & Papafragou, A. (2005). Language and thought. In R. Morrison & K. Holyoak (Eds.), *Cambridge handbook of thinking and reasoning* (pp. 633–661). Cambridge, England: Cambridge University Press.
- Gleitman, L., & Rozin, P. (1977). The structure and acquisition of reading I: Relations between orthographies and the structure of language. In A. Reber & D. Scarborough (Eds.), *Toward a psychology of reading* (pp. 447–493). Hillsdale, NJ: Erlbaum.
- Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago, IL: University of Chicago Press.
- Goldin-Meadow, S. (2003). Thought before language: Do we think ergative? In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 493–522). Cambridge, MA: MIT Press.
- Gordon, P. (2004). Numerical cognition without words: Evidence from Amazonia. *Science*, 306, 496–499.
- Goutteux, S., Thinus-Blanc, C., & Vauplair, S. (2001). Rhesus monkeys use geometric and nongeometric information during a reorientation task. *Journal of Experimental Psychology*, 130, 505–519.
- Grice, P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics, Vol. 3: Speech acts* (pp. 41–58). New York: Academic Press.
- Gumperz, J., & Levinson, S. (Eds.). (1996). *Rethinking linguistic relativity*. Cambridge, England: Cambridge University Press.
- Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour*, 61, 139–151.
- Heider, E., & Oliver, D. C. (1972). The structure of color space in naming and memory for two languages. *Cognitive Psychology*, 3, 337–354.
- Hermer, L., & Spelke, E. (1994). A geometric process for spatial representation in young children. *Nature*, 370, 57–59.
- Hermer, L., & Spelke, E. (1996). Modularity and development: The case of spatial reorientation. *Cognition*, 61, 195–232.
- Hermer-Vasquez, L., Spelke, E., & Katsnelson, A. (1999). Sources of flexibility in human cognition: Dual-task studies of space and language. *Cognitive Psychology*, 39, 3–36.
- Hespos, S., & Spelke, E. (2000). *Conceptual precursors to spatial language: Categories of containment*. Paper presented at the Meeting of the International Society on Infant Studies, Brighton, England.
- Hume, D. (1739/2000). *A treatise on human nature*. (D. F. Norton & M. Norton, Eds.). New York: Oxford University Press.
- Hurewitz, F., Papafragou, A., Gleitman, L., & Gelman, R. (2006). Asymmetries in the acquisition of numbers and quantifiers. *Language Learning and Development*, 2, 77–96.
- Imai, M., & Gentner, D. (1997). A crosslinguistic study of early word meaning: Universal ontology and linguistic influence. *Cognition*, 62, 169–200.
- Jackendoff, R. (1990). *Semantic structures*. Cambridge, MA: MIT Press.
- Jameson, D., & Hurvich, L.M. (1978). Dichromatic color language: "Reds" and "greens" do not look alike but their colors do. *Sensory Processes*, 2, 146–155.
- January, D., & Kako, E. (2007). Re-evaluating evidence for linguistic relativity. *Cognition*, 104, 417 – 426.
- Joshi, A. (1985). Tree adjoining grammars: How much context-sensitivity is necessary to provide reasonable structural descriptions? In D. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural language parsing* (pp. 206–250). Cambridge, England: Cambridge University Press.
- Jusczyk, P. (1985). On characterizing the development of speech perception. In J. Mehler & R. Fox (Eds.), *Neonate cognition: Beyond the blooming buzzing confusion* (pp. 199–229). Hillsdale, NJ: Erlbaum.
- Kay, P. (1996). Intra-speaker relativity. In J. Gumperz & S. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 97–114). Cambridge, England: Cambridge University Press.
- Kay, P., & Kempton, W. (1984). What is the Sapir-Whorf hypothesis? *American Anthropologist*, 86, 65–79.
- Kay, P., & Regier, T. (2002). Resolving the question of color naming universals. *Proceedings of the National Academy of Sciences USA*, 100(15), 9085–9089.
- Kay, P. & Regier, T. (2007). Color naming universals: The case of Berinmo. *Cognition*, 102, 289–298.
- Keenan, E., & Stavi, J. (1986). A semantic characterization of natural language determiners. *Linguistics and Philosophy*, 9, 253–326.
- Keller, H. (1955). *Teacher: Anne Sullivan Macy*. Westport, CT: Greenwood Press.

- Kellman, P. (1996). The origins of object perception. In R. Gelman & T. Au (Eds.), *Perceptual and cognitive development* (pp. 3–48). San Diego, CA: Academic Press.
- Kinzler, K. D., Shutts, K., DeJesus, J., & Spelke, E. S. (2009). Accent trumps race in guiding children's social preferences. *Social Cognition*, 27(4), 623–634.
- Krifka, M. (1995). Common nouns: A contrastive analysis of Chinese and English. In G. Carlson & F. J. Pelletier (Eds.), *The generic book* (pp. 398–411). Chicago, IL: University of Chicago Press.
- Kuhl, P., Williams, K., Lacerda, F., Stevens, K., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by six months of age. *Science*, 255, 606–608.
- Landau, B., Dessalegn, B., & A. Goldberg (2009). Language and space: Momentary interactions. In P. Chilton & V. Evans (Eds.), *Language, cognition, and space: The state of the art and new directions*. Advances in Cognitive Linguistics Series. London: Equinox Publishing.
- Landau, B., & Gleitman, L. (1985). *Language and experience: Evidence from the blind child*. Cambridge, MA: Harvard University Press.
- Landau, B., Smith, L., & Jones, S. (1998). The importance of shape in early lexical learning. *Cognitive Development*, 3, 299–321.
- Learmouth, A., Nadel, L., & Newcombe, N. (2002). Children's use of landmarks: Implications for modularity theory. *Psychological Science*, 13, 337–341.
- Leslie, A., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, 25, 265–288.
- Levelt, W. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
- Levinson, S. C. (1996). Frames of reference and Molyneux's question: Crosslinguistic evidence. In P. Bloom, M. Pederson, L. Nadel, & M. Garrett (Eds.), *Language and space* (pp. 109–169). Cambridge, MA: MIT Press.
- Levinson, S. C. (2003). Space in language and cognition: Explorations in linguistic diversity. Cambridge, England: Cambridge University Press.
- Li, P., Abarbanel, L., Gleitman, L., & Papafragou, A. (2011). Spatial reasoning in Tenejapan Mayans. *Cognition*, 120, 33–53.
- Li, P., Dunham, Y., & Carey, S. (2009). Of substance: The nature of language effects on entity construal. *Cognitive Psychology*, 58(4), 487–524.
- Li, P., & Gleitman, L. (2002). Turning the tables: Spatial language and spatial cognition. *Cognition*, 83, 265–294.
- Li, P., Gleitman, L., Landau, B., & Gleitman, H. (1997). Space for thought. *Paper presented at the 22nd Boston University Conference on Language Development*, Boston MA.
- Liberman, A. M. (1970). The grammars of speech and language. *Cognitive Psychology*, 1, 301–323.
- Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74, 431–461.
- Lidz, J., Gleitman, H., & Gleitman, L. (2003). Understanding how input matters: Verb learning and the footprint of universal grammar. *Cognition*, 87, 151–178.
- Lindsey, D. T., Brown, A. M., Reijnen, E., Rich, A. N., Kuzmova, Y. I., & Wolfe, J. M. (2010). Color channels, not color appearance or color categories, guide visual search for desaturated color targets. *Psychological Science*, 1208.
- Liu, Q., Li, H., Campos, J. L., Wang, Q., Zhang, Y., Qiu, J., Zhang, Q. L., & Sun, H.-J. (2009). The N2pc component in ERP and the lateralization effect of language on colour perception. *Neuroscience Letters*, 454, 58–61.
- Locke, J. (1690/1964). *An essay concerning human understanding*. (A. D. Woolley, Ed.). Cleveland, OH: Meridian Books.
- Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive Psychology*, 50, 315–353.
- Lucy, J. (1992). *Grammatical categories and cognition: A case study of the linguistic relativity hypothesis*. Cambridge, England: Cambridge University Press.
- Lucy, J., & Gaskins, S. (2001). Grammatical categories and the development of classification preferences: A comparative approach. In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 257–283). Cambridge, England: Cambridge University Press.
- Majid, A., Bowerman, M., Kita, S., Haun, D. B., & Levinson, S. C. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Science*, 8(3), 108–114.
- Malt, B., Sloman, S., & Gennari, S. (2003). Universality and language specificity in object naming. *Journal of Memory and Language*, 49, 20–42.
- Malt, B., Sloman, S., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, 40, 230–262.
- Malt, B., & Wolff, P. (Eds.). (2010). *Words and the mind: How words capture human experience*. Oxford, England: Oxford University Press.
- Mandler, J. (1996). Preverbal representation and language. In P. Bloom, M. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and space* (pp. 365–384). Cambridge, MA: MIT Press.
- Markman, E., & Hutchinson, J. (1984). Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive Psychology*, 16, 1–27.
- Mazuka, R., & Friedman, R. (2000). Linguistic relativity in Japanese and English: Is language the primary determinant in object classification? *Journal of East Asian Linguistics*, 9, 353–377.
- McDonough, L., Choi, S., & Mandler, J. M. (2003). Understanding spatial relations: Flexible infants, lexical adults. *Cognitive Psychology*, 46, 229–259.
- Mehler, J., & Nespor, M. (2004). Linguistic rhythm and the development of language. In A. Belletti & L. Rizzi (Eds.), *Structures and beyond: The cartography of syntactic structures* (pp. 213–221). Oxford, England: Oxford University Press.
- Munnich, E., Landau, B., & Dosher, B. A. (2001). Spatial language and spatial representation: A cross-linguistic comparison. *Cognition*, 81, 171–207.
- Naigles, L., & Terrazas, P. (1998). Motion-verb generalizations in English and Spanish: Influences of language and syntax. *Psychological Science*, 9, 363–369.
- Nunberg, G. (1978). *The pragmatics of reference*. Bloomington: Indiana University Linguistics Club.
- Ozturk, O., & Papafragou, A. (2008). The acquisition of evidentiality and source monitoring. In *Proceedings from the 32nd Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.
- Papafragou, A. (2007). Space and the language-cognition interface. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind: Foundations and the future* (pp. 272–292). Oxford, England: Oxford University Press.
- Papafragou, A., Hulbert, J., & Trueswell, J. (2008). Does language guide event perception? Evidence from eye movements. *Cognition*, 108, 155–184.

- Papafragou, A., Li, P., Choi, Y., & Han, C. (2007). Evidentiality and the language/cognition interface. *Cognition*, 103, 253–299.
- Papafragou, A., Massey, C., & Gleitman, L. (2002). Shake, rattle ‘n’ roll: The representation of motion in language and cognition. *Cognition*, 84, 189–219.
- Papafragou, A., Massey, C., & Gleitman, L. (2006). When English proposes what Greek presupposes: The cross-linguistic encoding of motion events. *Cognition*, 98, B75–87.
- Papafragou, A., & Musolino, J. (2003). Scalar implicatures: Experiments at the semantics-pragmatics interface. *Cognition*, 86, 153–182.
- Papafragou, A., & Selimis, S. (2010a). Lexical and structural biases in the acquisition of motion verbs. *Language Learning and Development*, 6, 87–115.
- Papafragou, A., & Selimis, S. (2010b). Event categorisation and language: A cross-linguistic study of motion. *Language and Cognitive Processes*, 25, 224–260.
- Pederson, E., Danziger, E., Wilkins, D., Levinson, S., Kita, S., & Senft, G. (1998). Semantic typology and spatial conceptualization. *Language*, 74, 557–589.
- Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.
- Prasada, S., Ferenz, K., & Haskell, T. (2002). Conceiving of entities as objects and stuff. *Cognition*, 83, 141–165.
- Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.
- Quinn, P. (2003). Concepts are not just for objects: Categorization of spatial relational information by infants. In D. Rakison & L. Oakes (Eds.), *Early category and object development: Making sense of the blooming, buzzing confusion* (pp. 50–76). Oxford, England: Oxford University Press.
- Ratliff, K., & Newcombe, N. (2008). Is language necessary for human spatial reorientation? Reconsidering evidence from dual task paradigms. *Cognitive Psychology*, 56(2), 142–163.
- Regier, T., Kay, O., & Cook, R. S. (2005). Focal colors are universal after all. *Proceedings of the National Academy of Sciences USA*, 102, 8386–8391.
- Regier, T., Kay, P., Gilbert, A., & Ivry, R. (2010). Language and thought: Which side are you on, anyway? In B. Malt & P. Wolff (Eds.), *Words and the mind: How words capture human experience* (pp. 165–182). New York: Oxford University Press.
- Regier, T., Kay, P., & Khetarpal, N. (2009). Color naming and the shape of color space. *Language*, 85, 884–892.
- Restle, F. (1957). Discrimination of cues in mazes: A resolution of the place-vs.-response question. *Psychological Review*, 64, 217–228.
- Roberson, D. (2005). Color categories are culturally diverse in cognition as well as in language. *Cross-Cultural Research*, 39, 56–71.
- Roberson, D., Davidoff, J., Davies, I., & Shapiro, L. (2005). Colour categories in Himba: Evidence for the cultural relativity hypothesis. *Cognitive Psychology*, 50, 378–411.
- Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, 129, 369–398.
- Sapir, E. (1941). In L. Spier (Ed.), *Language, culture and personality: Essays in memory of Edward Sapir*. Menasha, WI: Memorial Publication Fund. Cited in Whorf (1956, p. 134).
- Senghas, A., Coppola, M., Newport, E., & Suppala, T. (1997). Argument structure in Nicaraguan Sign Language: The emergence of grammatical devices. In *Proceedings of the 21st Annual Boston University Conference on Language Development* (pp. 550–561). Somerville, MA: Cascadilla Press.
- Slobin, D. (1996). From ‘thought and language’ to ‘thinking for speaking’. In J. Gumperz & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 70–96). Cambridge, England: Cambridge University Press.
- Slobin, D. (2001). Form-function relations: How do children find out what they are? In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 406–449). Cambridge, England: Cambridge University Press.
- Smith, L. (2001). How domain-general processes may create domain-specific biases. In M. Bowerman & S. C. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 101–131). Cambridge, England: Cambridge University Press.
- Smith, L., Colunga, E., & Yoshida, (2001). Making an ontology: Cross-linguistic evidence. In D. Rakison & L. Oakes (Eds.), *Early category and object development: Making sense of the blooming, buzzing confusion* (pp. 275–302). Oxford, England: Oxford University Press.
- Snedeker, J., Thorpe, K., & Trueswell, J. (2001). On choosing the parse with the scene: The role of visual context and verb bias in ambiguity resolution. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.
- Soja, N., Carey, S., & Spelke, E. (1991). Ontological categories guide young children’s inductions of word meaning: Object terms and substance terms. *Cognition*, 38, 179–211.
- Spelke, E., & Tsivkin, S. (2001a). Language and number: A bilingual training study. *Cognition*, 78, 45–88.
- Spelke, E., & Tsivkin, S. (2001b). Initial knowledge and conceptual change: Space and number. In M. Bowerman & S. C. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 70–100). Cambridge, England: Cambridge University Press.
- Spelke, E., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). The origins of knowledge. *Psychological Review*, 99, 605–632.
- Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Cambridge, MA: Harvard University Press.
- Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), *Language typology and syntactic description* (pp. 57–149). New York: Cambridge University Press.
- Tanenhaus, M. K. (2007). Spoken language comprehension: insights from eye movements. In G. Gaskell (Ed.), *Oxford handbook of psycholinguistics* (pp. 309–326). Oxford, England: Oxford University Press.
- Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition*, 74, 209–253.
- Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology*, 14(1), 107–141.
- Trueswell, J., & Papafragou, A. (2010). Perceiving and remembering events cross-linguistically: Evidence from dual-task paradigms. *Journal of Memory and Language*, 63, 64–82.
- Trueswell, J. C., Sekerina, I., Hill, N. M., & Logrip, M. L. (1999). The kindergarten path effect: Studying on-line sentence processing in young children. *Cognition*, 73, 89–134.
- Vallortigara, G., Zanforlin, M., & Pasti, G. (1990). Geometric modules in animals’ spatial representations: A test with chicks. *Journal of Comparative Psychology*, 104, 248–254.
- Varley, R., & Siegal, M. (2000). Evidence for cognition without grammar from causal reasoning and ‘theory of mind’ in an agrammatic aphasic patient. *Current Biology*, 10, 723–726.

- Welder, A. N., & Graham, S. A. (2001). The influence of shape similarity and shared labels on infants' inductive inferences about nonobvious object properties. *Child Development*, 72, 1653–1673.
- Werker, J. F., & Lalonde, C. E. (1988). Cross-language speech perception: Initial capabilities and developmental change. *Developmental Psychology*, 24(5), 672–683.
- Werker, J., & Logan, J. (1985). Cross-language evidence for three factors in speech perception. *Perception and Psychophysics*, 37, 35–44.
- Werker, J., & Tees, R. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7, 49–63.
- Werker, J. F., & Desjardins, R. N. (1995). Listening to speech in the first year of life: Experiential influences on phoneme perception. *Current Directions in Psychological Sciences*, 4(3), 76–81.
- Whorf, B. L. (1956). Language, thought and reality (J. Carroll, Ed.). Cambridge, MA: MIT Press.
- Winawer, J., Witthoft, N., Frank, M. C., Wu, L., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences USA*, 104(19), 7780–7785.
- Wittgenstein, L. (1922/1961). *Tractatus Logico-Philosophicus* (D. F. Pears & B. McGuinness, Trans.). London: Routledge.
- Xu, F. (2002). The role of language in acquiring object kind concepts. *Cognition*, 85, 223–250.

Thinking in Societies and Cultures

Tage S. Rai

Abstract

Rather than focusing on cross-cultural differences in cognition, this chapter focuses on how being born into and living in any large sociocultural group is deeply intertwined with thinking. I begin by discussing the ways in which thinking in societies and cultures is heavily distributed across people, time, and technology. This selective review focuses primarily on research on distributed and extended cognition, social and transactive memory, and cognitive anthropology and cultural transmission. The second part of the chapter addresses the question of whether cultural evolution is linked to cognitive evolution in humans. In particular, I will ask to what extent our minds are predisposed to acquire knowledge relevant for succeeding in culture, and how this possibility changes our view of the role of culture in cognition. I will conclude the chapter by proposing a division between micropsychological and macropsychological levels of analysis in psychology, and I will suggest research questions that this perspective is uniquely suited to address.

Key Words: macropsychology, distributed cognition, cognitive anthropology, cultural transmission, cultural evolution, social biases, social learning, enculturation, group psychology, social media

Introduction

Without men, no culture, certainly; but equally, and more significantly, without culture, no men. (Geertz, 1973, p. 49)

Although it may be unintended, the disciplinary term “culture and cognition” implies a clean distinction between minds and the sociocultural contexts within which minds exist and must operate. Similarly, reviews of culture and cognition typically focus on *cross-cultural differences* in thinking, such as the differences in attention, categorization, causal attribution, or causal reasoning reported between individualistic (independent, analytic) cultures and collectivistic (interdependent, holistic) cultures (for recent comprehensive reviews of this literature, see Heine & Ruby, 2010; Lehman, Chiu, & Schaller, 2004; Markus & Kitayama,

1991; Nisbett, 2004; Nisbett & Norenzayan, 2002; Nisbett, Peng, Choi, & Norenzayan, 2001; Segall, Lonner, & Berry, 1998; Triandis, 1989). In this chapter, I would like to question the distinction between culture and cognition, while considering basic and often overlooked ways in which our thinking is inextricably linked to life in societies and cultures (see Fessler & Machery, in press; Fiske et al., 1998).

Just as there are interactions between the cell and the individual mind, there are interactions between the individual mind and the social group, such that the cognitive processes of both are influenced by the existence of the other (Brown, Collins, & Duguid, 1989; Hutchins, 1995a; Lave & Wenger, 1991). What I term “macropsychology”—the complexity of human cognition that emerges in sociocultural groups—concerns how living and being born into

any human culture changes the way we think, what we think about, and how our minds have evolved.

For our purposes, I will define human societies simply as large groups of socially interacting individuals. Notably, modern societies are often characterized by high levels of access to technology, which as I address later, has significant downstream effects on cognition. Although historically contested (Lug-hod, 1991), I will define culture simply as the set of shared meanings, knowledge, and practices that are transmitted within societies across generations. Culture is embodied in a society's objects, tools, language, artifacts, rituals, technologies, and media. Note that the entire body of culture is not shared by everyone within a society. Issues of communication and physical access will prevent everyone from participating in all aspects of culture, while issues of trust and legitimization will create cultural barriers to sharing information between subgroups within a society. As far as a definition of "thinking," I will focus on how cognition within societies and cultures is characterized by "thinking for doing" (Fiske, 1992), wherein processes such as memory, reasoning, and decision making occur in the service of navigating social-relational interactions.

Rather than providing an exhaustive review of all of the ways in which thought and culture are intertwined, my goal in this chapter is to map out the research terrain so as to highlight key areas for future study, and to spur research on the intersection of thought and culture. In particular, I will draw on relevant research in cognitive anthropology and theories of distributed cognition, social neuroscience and theories of cultural evolution and transmission, and sociological and psychological approaches to cognition in social groups. Much of this work is captured under similar headings, such as distributed, situated, extended, and embodied cognition; or social, collective, transactive, and collaborative memory. Although relevant for thinking about group cognition, I will not focus in great detail on the vast literature documenting social biases in individual cognition, such as social influences on eyewitness memory (Loftus & Hoffman, 1989) or the effects of stereotypes on performance (Steele & Aronson, 1995). Instead, I will only consider social biases that capture emergent properties at the level of the group, such as research on group polarization or conformity. I will also not focus on proposed instances of culture or tool use in other mammals, as it is simply not on the same scale as that of humans (see Penn & Povinelli, Chapter 27). Nor will I

consider in detail research on artificial intelligence, or the swarm intelligence evident in colonies of ants, schools of fish, or flocks of birds, because my focus in this chapter is restricted to human thinking and how it has been affected by social life. Through this selective review, I hope to shed light on how the advent of society connected through culture generates macropsychology by distributing thinking across people, time, and technology.

By reviewing the distributed nature of cognitive processes such as thinking, knowing, reasoning, learning, remembering, and decision making, it will become apparent that living in societies connected by culture is not only necessary for acquiring specific kinds of knowledge (e.g., how to perform an initiation rite or how to dress for the festival), but that it is crucial for *producing* human cognition in arguably its most important respects. Specifically, culturally produced cognition enables us to build on and recombine the ideas of others by distributing cognitive processes, generating collective representations, and selecting for cognitive abilities that are advantageous to individuals living in social groups. At the same time, conditions or events that affect the social group will have major implications for individual-level thinking, reasoning, knowledge, and preferences. With this perspective in mind, we can ask two sets of interrelated questions.

1) How does participation in societies and cultures facilitate a division of cognitive labor across people, time, and technology? How is information represented and transmitted in this system and what are its consequences?

2) How has individual thinking adapted specifically for learning and succeeding in a sociocultural setting, both in terms of brain evolution and the cognitive capacities we are predisposed to develop? Would the removal of culture alter cognitive processes or the informational content informing those processes, or both? Is such a hypothetical state even possible?

To examine these questions, I will consider the nature of individual knowledge and how we leverage the expertise of others, how knowledge is accumulated and transmitted via social learning and cultural evolution, the sensitivity of our thinking to social sources of information and the impact of group changes on our thinking, how technology augments our memory capacity and processing capabilities, and the ways in which our minds may be specifically adapted for learning culture.

I will conclude the chapter by discussing how a fuller understanding of the relationship between culture and thought can transform the nature of psychological inquiry and give rise to quite promising new avenues of research into both individual and group psychology. In particular, I will suggest that we must examine the emergent macropsychology of social groups, as cognitive processes cannot be reduced to the aggregate of individual psychologies. At the same time, we will have to bolster studies of individual psychology with an understanding of the socio-cultural contexts in which it developed. Within this framework, observational and descriptive research outside of the laboratory will be required in order to inductively develop and test theory against real-world behavior. Guiding questions will include how individual thinking is embedded in culture, how we should represent cognitive processes at the level of the group, in what ways we may be specifically attuned to learning and thinking with others, and the ways in which thinking sometimes occurs independent of knowledge.

Cognition Through Culture

For socially marginalized, otherwise quite intelligent adolescent boys (i.e., “nerds”), there is no better pastime than imagining how great life would be if they could travel back in time. Just as the Connecticut Yankee awoke in King Arthur’s court following a blow to the head (Twain, 1889), they too would reemerge from the lockers they were stuffed in to find a preindustrial Camelot that they could rule with their modern knowledge and superior intellect. Unfortunately, for the nerds, what they fail to realize is that they—and people in general—actually don’t know how to do much of anything *completely on their own*. Rather, a given individual, no matter how intelligent, knows incredibly little.

Thinking Distributed Across People

Even an expert on modern firearms, such as the Connecticut Yankee, would be helpless without any idea as to how to produce necessary raw materials, such as lead or gunpowder, from scratch. Similarly, I know how to turn on a switch, but I don’t have any deep understanding of how my television or my microwave works, let alone the myriad causal processes that interact between my home, a power plant, and the earth’s magnetic field (is this even involved?) to generate electricity.

What about a relatively simple device, such as a zipper? In any deep sense, do you have any idea how

exactly a zipper “zips”? In a series of experiments, Keil (2003, 2006, 2010; Rozenblit & Keil, 2002) has documented our lack of explanatory understanding. In one study with *graduate students at Yale*, participants were asked to write detailed explanations for everyday artifacts. It was found that explanations of even simple objects such as zippers or piano keys were quite shallow. At the same time, their confidence in their explanatory knowledge was extraordinarily high (Rozenblit & Keil, 2002).²

If we understand so little and fail to realize it, how do we succeed? Keil (2003, 2006) has argued that we overcome our shallow understanding by successfully leveraging the knowledge of other people. Such a situation is familiar to scientists, who often succeed not by knowing everything themselves, but by knowing who to consult. In support of this view, Keil (2010) has found that children as young as 8 years of age can distinguish categories on the basis of their likelihood to have experts (e.g., hammers used by miners vs. hammers with red handles).

Wegner (1987; also see Hollingshead, 1998) has found that romantic couples leverage each other’s expertise by dividing the work of remembering between each other, a process he refers to as “transactive memory.” When tested together, the extent to which participants will remember new information depends in part on the expertise *their partner* has in the knowledge domain. Transactive memory systems have also been investigated in the context of production in work teams in both short-term (Liang, Moreland, & Argote, 1995) and long-term tasks (Austin, 2003). It has been found that training groups together on a task (e.g., building a transistor radio) can increase memory for the procedure and subsequent performance at test. Such transactive memory processes likely include encoding of information as well. When a member of your transactive memory system encounters information of which they know it would be important for you to be aware, they can encode that information and transmit it to you at a later time. They can even encode the source of that information so that you can engage in deeper level processing of the information on your own (Wegner, 1987).

In contrast, studies of “collaborative memory” have typically argued that remembering in groups inhibits memory function. These studies have shown that although social groups working together do recall more than a single individual, they recall less than a group of individuals working separately (Weldon & Bellinger, 1997). However, there are

two causes for possible concern with this line of inquiry. First, it disregards the fact that “pooling” the responses of individuals working separately is itself social in nature, requiring a hierarchical distribution of responsibilities that may occur in many real-life collaborative efforts. Second, although groups of individuals working separately on a single small task may outperform groups working together, groups working together will have the ability to distribute their cognitive efforts across several large tasks and build on the efforts of each other over time to accumulate and transmit a much greater shared body of knowledge.

Representation and Transmission Across Time

Living in social groups greatly enhances the cognitive capacity of a given individual because we can rely on others for both additional memory and information processing, without having to learn or remember nearly so much on our own. Information transmitted from others provides a sort of scaffolding upon which individual learning can build. In turn, individual learning can lead to modifications in previous knowledge, behaviors, and practices. Thus, our modern knowledge actually reflects an accumulation of information socially transmitted across generations (Henrich & McElreath, 2003; Richerson & Boyd, 2005; Tomasello, 1999).

To build on the contributions of other people and to consolidate their knowledge into chunks of information that we can understand, we need culture to *mediate* our access to this shared body of knowledge (Cole, 1995; Vygotsky, 1978). Drawing on schema theories, D’Andrade (1981) argued that we employ cultural schemas to facilitate use of this shared body of knowledge by organizing the information into usable chunks or scripts imbued with meaning, such that access to one part of the cultural model and its meaning (e.g., walking into a restaurant) will activate the rest of the cultural model (e.g., wait to be seated, open menu, and order food). Importantly, because of the distributed nature of thinking, individuals involved in a situation will never have access to all of the information, but they will share enough knowledge to coordinate, much as in Keil’s (2010) example of expert knowledge. Thus, I may have no knowledge of what a waiter does with my order, but both the waiter and I know what our roles and responsibilities are.

The mechanisms by which people in a social group acquire culturally mediated knowledge from

others are varied, but perhaps the most important is learning through observation and subsequent imitation, processes by which individuals actively attempt to reproduce behaviors they have seen (Lancy, 1996; Tomasello, 1999). The classic study of observational and imitative modeling was Bandura, Ross, and Ross’s (1963) experiment in which they found that children were more likely to engage in aggressive behaviors toward a doll if they had previously observed an adult engaging in such behaviors. In naturalistic settings, observational and imitation-based learning is evident in the guided participation characteristic of apprenticeship in non-Western cultures (Brown et al., 1989; Lave & Wenger, 1991). In these situations, there is little explicit teaching by masters. Rather, apprentices observe masters, imitate their actions without complete understanding, and slowly gain mastery themselves as they take on greater responsibilities over time (e.g., when Daniel-san unknowingly learns karate by helping Mr. Miagi fix up his house).

Experimental investigations into how these learned behaviors are transmitted and modified over time have focused on the development of strategies within “micro-societies” (Caldwell & Millen, 2008). In these studies, participants are removed and replaced from groups as they try to design basic artifacts in order to simulate generational change. For example, in one study groups of participants were tasked with constructing paper airplanes that could travel as far as possible. At any given time, some participants were constructing airplanes while others observed and asked questions of them. Over time, those who constructed their airplanes tested them and were subsequently replaced by those who observed them, with new observers entering the scene. Completed airplanes were kept available with their recorded distances for all subsequent group members, such that all participants had a historical record of all previous attempts. This process was meant to simulate vertical transmission across generations, in which a given learner has access to the knowledge and efforts of previous generations.

Caldwell and Millen found that within a 10-person group, the performance of airplanes significantly improved across generations: on average, the paper airplane produced by the tenth participant traveled nearly three times further than the airplane produced by the first participant. The similarity of airplane designs within a group, as assessed by a separate group of naïve raters, was closer between airplanes of successive generations than between generations

that were farther apart, as would be expected if transmission with gradual modification was occurring. Similarity was also closer among airplanes from the same group than between airplanes from different 10-person groups, suggesting a process analogous to cultural variation took place. However, similarity between different groups increased over time; paper airplanes made by the eighth, ninth, and tenth participants in the groups were more similar to their counterparts in the other groups than those made by earlier participants, suggesting a process of convergent evolution toward objectively more efficient designs (Caldwell & Millen, 2008; also see Griffiths, Kalish, & Lewandowsky, 2008).

Culturally mediated social learning via observation, imitation, or other mechanisms generally makes sense when it is difficult to learn things on your own.³ Moreover, we do not copy just anyone, but rather we are biased toward learning from *particular* others. For example, children are often biased toward horizontal imitation of same-age or slightly older children rather than vertical imitation of parents, likely because other children are the models from whom they would have the most to learn in many cases (Harris, 1995). Similarly, we may be biased toward copying the successful and conforming to the majority. Copying successful models rather than learning yourself is generally advantageous when individual learning is difficult, and when there is a large amount of variation in the abilities between the successful and the unsuccessful (Henrich & Gil-White, 2001). Success-biased transmission was evident in my (failed) attempts to copy and incorporate Michael Jordan's fade-away jumper into my basketball game as a child. Similarly, conforming to the majority may be more advantageous than individual learning when learning is difficult because conformity is roughly equivalent to basing actions on a large sample (Henrich & Boyd, 1998). Conformity-biased transmission is evident in odd fashion trends, such as my regrettable but effective attempts to copy the use of flannel shirts during the early 1990s. Finally, it is likely that we also biased toward selectively copying people to whom we are connected by social relationships, such as friends and family members as opposed to strangers (Fiske, 1991).

Coconstruction, Sampling, and Isolation

COCONSTRUCTION OF KNOWLEDGE

With the advent of language, processes of observation and imitation have been augmented by the learning of oral narratives. In such situations, knowledge

is reproduced and modified with each telling to facilitate even greater maintenance and accumulation of knowledge over time (Boyer, 1992; Sperber, 1985). In modern contexts, such oral transmission is codified in the coconstruction of history; our collective memory of the past (Halbwachs, 1950; Olick & Robbins, 1998; Zerubavel, 1996). Specifically, history is coconstructed by recombining the individual memories of large groups of people to form a narrative from which later generations may learn. Due to our lack of direct access to the past events recounted in the narrative, history is continually reconstructed in response to the needs of the present (Schwartz, 1991). For example, Welzer (2010) has found that when German grandchildren interview their grandparents about World War II, they actively change various aspects of their grandparents' narrative to satisfy their own psychological need to deal with a collective trauma. These alterations are mutually supported by members of the younger generation and transmitted horizontally within the younger generation to reconstruct history into a new, modified version of itself. Because of this coconstructive nature of history, it cannot be reduced to individual memory, and it must instead be considered a collective representation (Halbwachs, 1950)—an invention made possible by individuals living together in culturally connected societies.

SAMPLING BIASES

One basic consequence of learning so much from other people is that many of our beliefs and attitudes are largely a product of where we grew up. Latane (1981, 1996) has argued that because people close to you will naturally have more impact on your learning, shared beliefs and attitudes will begin to cluster in physical space, and sets of different beliefs will begin to correlate with each other, in a process he refers to as "dynamic social impact." Latane and Bourgeois (1996) have demonstrated that differential sampling of beliefs leads to clustering of beliefs among the groups that are sampled, correlations among different beliefs within these groups, consolidation of minority groups, and some level of diversity even when everyone attempts to be in the majority.

The potentially adverse impact of relying on sampling rather than individual learning is evident in studies of information cascades (Anderson & Holt, 1997). In an information cascade, if the first few individuals make a particular judgment or express a particular belief, subsequent individuals will copy

their actions even if doing so contradicts their own beliefs. Specifically, if my own calculation supports hypothesis A, but the *expressed* judgments of everyone before me support competing hypothesis B, then I will weigh their expressed judgments as evidence for competing hypothesis B and may express a judgment in favor of hypothesis B in spite of my own calculations. Such a process can generate a cascade of conformity in which all subsequent individuals support the competing hypothesis, even if their own individual calculations support a different hypothesis (also see Asch, 1955).

The errors that can result from misrepresentations of other actors' internal beliefs are also evident in research on group polarization (Isenberg, 1986; Moscovici & Zavalloni, 1969) and pluralistic ignorance (Miller & Prentice, 1994). Group polarization refers to the tendency of group decisions to be more extreme than the average decisions that would have been made by individuals. For example, in jury decision making, awards of damages following group discussion become either significantly more severe or more lenient than the average decisions made by individuals (Bray & Noble, 1978). In cases of pluralistic ignorance, members of a group express public views that differ from their inner views. They fail to express their inner views because they assume that everyone else's inner views correspond to their public views and worry about the social costs of publicly dissenting. For example, Miller and Prentice argued that pluralistic ignorance can lead members of juvenile gangs who did not harbor extreme antisocial attitudes when they first joined to "act tough" to match their perceptions of everyone else's positions. As a consequence of this process, all the gang members may pursue antisocial behavior that none of them would have preferred individually.

Inventions of modern societies, such as prediction games and markets, can offset some of the risks of information cascades, pluralistic ignorance, and group polarization. In prediction games and markets, large groups of individuals make predictions about the likelihood of events ranging from sports victories to presidential winners (Wolfers & Zitzewitz, 2004). It has been argued that these large groups of individuals making predictions are more accurate in the aggregate than any expert due to the inherent advantages of large samples over small ones (Surowiecki, 2005). The key is for individuals to make their predictions independently so as to avoid the problem of information cascade, and to make their decisions anonymously so as to avoid the social

costs that lead to pluralistic ignorance. Surowiecki argued that when prediction markets have failed, such as in the case of stock market bubbles, it has been because the conditions of independence and anonymity were not satisfied. As I will touch on later, the ability to draw on the wisdom of crowds in making judgments that are superior to individuals requires culturally mediated information technologies that are themselves products of living in modern societies.

COSTS OF ISOLATION

Perhaps more intriguing than the cognitive consequences of participating in a social group are the cognitive consequences of being cut off from the group. Wegner (1987) hypothesized that the cognitive impairments reported by recent divorcees may not be the result of psychological trauma, but rather may actually reflect the loss of an integral part of a transactive memory system. Specifically, if romantic couples naturally and implicitly divide the work of remembering between them, it may take time for former partners to reallocate resources to areas that were once outside of their cognitive responsibilities. Similarly, analyses of poverty have argued that low-socioeconomic status (SES) individuals lack access to the extensive environmental and interpersonal supports that people in high-SES environments rely on for everyday cognition. As a consequence, living poor in America is akin to a chronic state of thinking under divided attention, a condition known to severely impair cognitive processing (Bertrand, Mullainathan, & Shafir, 2004).

Historical examinations of rapid population change have also demonstrated the consequences of changes to the group on cognition. For instance, when discovered by Europeans, Tasmanians only employed about 24 simple tools with regularity. In contrast, the aboriginal Australians 200 kilometers away, who used to be connected to Tasmania via a land bridge, used hundreds more tools that were much more complex, including specialized spears and nets for hunting and trapping prey. Moreover, the archaeological record suggests that tools in Tasmania were more complex 10,000 years earlier (just before the land bridge to Australia was engulfed by the ocean) than at the time of initial European contact (Henrich, 2004). Henrich cites these changes as evidence that tools did not merely stop developing once the island became isolated from mainland Australia. Rather, it seems that a culturally accumulated body of knowledge related to tool use was gradually

lost once the number of individuals over whom knowledge could be distributed and from which it could be learned was greatly reduced.

Thoughts Extended Into Technology

Would you still like to know how a zipper works? Go to <http://science.howstuffworks.com/innovation/everyday-innovations/zipper.htm>.

Rozenblit and Keil (2002) interpreted their participants' overconfidence as an error in the participants' metacognitive awareness of their own knowledge given the shallow explanations they were able to provide in the experiment. However, their overconfidence may indicate that rather than stating the actual explanatory knowledge available to them *in their heads*, the participants were in fact reporting whether they would be able to find the answer through a process of search involving various social technologies. As the aforementioned Internet link exemplifies, much of our available knowledge is actually not inside any individual's head; rather, it is now recorded in various culturally mediated written, electronic, or oral media that participants can easily access.

Importantly, this fact raises an interesting question as to whether Rozenblit and Keil's (2002) participants actually made a metacognitive error, and by extension, whether they actually had shallow explanatory knowledge. If we do not define knowledge as restricted to the memory storage of an individual mind, but instead include external culturally mediated memory retrieval such as going to a library, surfing the Internet, or consulting experts who can perform their own searches, then participants' explanatory understandings may have been quite rich. Similarly, should we consider a person who has corrective glasses to have poor vision, or a person with a hearing aid to have poor hearing? From a practical standpoint, does it make sense to test students' ability to hold information in their head, or should exams be "open-book" in the same way that students are now allowed to use calculators on most mathematics tests?

In their work on the extended mind hypothesis, Clark and Chalmers (1998) have emphasized this broader conception of cognition, which includes artifacts and objects in the world that can unite the knowledge of large groups of individuals. Inventions such as writing and computers have enabled us to massively augment our long-term memory as well as facilitate our working memory, as in "carrying" processes during mathematical operations,

spatial tasks requiring rotation, or writing out our thoughts during brainstorming. For example, Clark and Chalmers argued that if two people were trying to get somewhere and one memorized the location while the other wrote it down in a notebook always on his or her person, both could be said to be utilizing their memories and have merely stored the information in different locations. External cognitive processes are often faster than cognitive processes inside the head. For example, in the game "Tetris," pressing the rotate button repeatedly until the correct configuration appears may be faster than mentally rotating the object yourself (Kirsh & Maglio, 1994). According to Clark and Chalmers, if we redefine cognition as extended into environmental supports, then we can actually think of processes such as writing as alternative forms of thinking. (For other reviews of embodied and extended cognition, see Anderson, 2003; Wilson, 2005.)

In a similar vein, Hutchins (1995b) argued that the distribution of cognitive tasks across individuals and objects, and the specialization of work it enables, is critical to the success of complex cognitive tasks such as landing a plane. For example, pre-calculated informational readouts ("speed cards") detail the speed a plane should travel given its gross weight in order to land successfully. According to Hutchins, speed cards and other similar technologies should not be thought of as long-term memory aids, because people cannot change the information on the speed card and because its existence enables pilots to engage in different cognitive processes, such as spatial cognition tasks. Rather, the speed card is the long-term memory for the cognitive system of the plane and its crew, of which an individual pilot is merely one part (also see Hutchins, 1995a).

Recently, social-collaborative technologies such as Wikipedia have distributed cognitive processing in ways never thought possible. Kittur and Kraut (2008) have argued that whereas before there were significant costs and losses of information in the process of collaboration, Wikipedia has reduced these costs through its ability to maintain fidelity of the knowledge built across individuals, allowing them to build on each other's efforts.⁴ The up side of such collaboration is that the quality of work written and rewritten by several people is uniformly higher than the efforts of a single individual. Thus, the level of quality of Wikipedia articles is equivalent to traditional encyclopedias, and the highest quality Wikipedia articles have had the most editors. This process of having large groups edit and reedit each

other's work can be considered a form of group-level thinking in which rough ideas are gradually modified into final forms.

In this framework, culturally mediated technologies have enabled the individual mind to become part of ever-expanding cognitive networks that solve complex problems no individual mind could solve on its own (Hutchins, 1995a). Perhaps most intriguing are cases where we have actually outsourced our thinking to technology. For example, the "genius" feature in Apple's iTunes takes pre-established preferences and uses them to generate new preferences. Via access to the preferences of a large group of individuals, iTunes can compute correlations among preferences for different types of music. With this information, it can form categories and subsequently perform inferences to identify new music for us (Brooks, 2007). It should be noted that societies and cultures vary considerably in their access to information technology. As touched upon earlier in the context of poverty, some cultures cannot afford to provide access, others restrict it, and others do not trust the people in control of it. Do such cultures know less? To what extent do some cultures overly rely on technology, thus being exposed to misinformation?

Cognition for Culture

The result of the culturally mediated distribution of thinking across people, time, and technology, is that a given individual *never needs to know*. Rather, the "cognitive division of labor" (Wilson & Keil, 1998) has changed cognition at the individual level by enabling individuals to focus their attention on specialized tasks that can complement others in the service of group goals. In the following sections, I will consider the possibility that living in societies connected by culture may have actually altered the evolution of our minds and the cognitive processes they perform.

Evolution of Cultural Cognition

If thinking is incredibly interconnected with social living, then it may be possible that our minds have coevolved with the emergence of large social group living (Richerson & Boyd, 2005; Tomasello, 1999). In terms of the evolution of neural systems, we know that in primates as compared to other species, the neocortex has grown disproportionately to the rest of the brain (Dunbar, 1993; see Morrison & Knowlton, Chapter 6). Importantly, the energetic costs of building larger brains likely required a greater

proportion of meat and fat in our diet, which in turn was likely made possible by socioculturally mediated inventions such as cooking and improvements in hunting (Richerson, Boyd, & Henrich, 2010).

It has been hypothesized that selection for changes to neocortex occurred in part to support social learning and the acquisition of culturally transmitted information (Henrich & McElreath, 2003). Support for this view is based on data indicating that across primates, neocortex size relative to body size is correlated with many aspects of complex social group living, including the size of the social group (Dunbar, 1992) and the propensity for social learning (Reader & Laland, 2002; for a review, see Dunbar & Shultz, 2007).

In terms of particular neural systems, functional imaging studies have revealed that dorsomedial prefrontal cortex and the medial parietal cortex are tonically deactivated by engaging in general cognitive and reasoning tasks when compared to the resting baseline activation, as opposed to the baseline activation established by a control condition.⁵ In contrast, these areas are activated beyond the resting baseline when participants view social-relational interactions, suggesting both that these areas function to process social relations, and that thinking about social relationships is what we do most of the time (Iacoboni et al., 2004). Similarly, one key to succeeding in social groups may be the ability to infer the inner mental states of others when observing their behavior. This process, referred to as "mindreading," may also support our capacities for observational and imitation-based learning (Gallese & Goldman, 1998; Rizzolatti & Craighero, 2004; but see Lillard, 1998).

In primates, it has been found that mirror neurons located primarily in the ventral premotor area fire both when performing a goal-directed action (e.g., grasping an object) and when observing another individual performing the same action. Importantly, these neurons will not fire if the action of the observed individual is not goal directed, for example, if the person is grasping at nothing (Rizzolatti & Craighero, 2004). Iacoboni et al. (2005) extended this work to humans by demonstrating that mirror neuron systems uniquely fired when participants viewed scenes for which differential interpretations of the actor's intentions were possible. It has been hypothesized that the mirror neuron system in humans has adapted beyond that of other primates to form the basis for a sophisticated simulation-based mindreading mechanism, necessary for successful social interaction (Gallese & Goldman, 1998).

Social-Functional Aspects of Thought and Predispositions to Learn Culture

If our minds have coevolved with culture, it is possible that some of our most basic cognitive functions have adapted specifically for succeeding in social groups, and that we are predisposed to acquire socially relevant skills without extensive learning. In the broader cognitive framework sketched in this chapter, natural selection can be understood as “learning encoded on a long timescale,” because it can select for early acquisition of particular psychological tendencies and preferences over time if there is stability in the relevant behavioral contexts across generations.

For example, if skills such as imitation or mind-reading are recurrently adaptive, then there may be selective pressure toward evolving predispositions for early acquisition of these skills. Meltzoff and Moore (1977) found that young infants can mimic the mouth gestures of people they see, suggesting an evolved predisposition to imitate other people. Gergely, Bekkering, and Kiraly (2002) have found that as early as 14 months of age, infants appear to be aware of the goal-directed nature of actions. When 14-month-old infants observed a model turn off a light with her head when her hands were free, the infants were likely to imitate the model’s head action in turning off the light. However, infants who observed the model using her head when her hands were occupied were significantly less likely to imitate this head action, suggesting that the infants had inferred that the only reason the model used her head was because her hands were occupied. Similarly, Onishi and Baillargeon (2005) have demonstrated that even 15-month-old infants can pass a nonverbal version of the false belief task, in which participants must infer whether another actor knows the location of a toy.

More extreme social-functional theories have argued that language evolved as a consequence of living in larger social groups, which required mechanisms for tracking the histories of the numerous social relationships that developed (Dunbar, 1996); or that reasoning evolved as an argumentative tool to convince others to support our positions (Mercier & Sperber, 2010). Although direct evidence for these theories is lacking, there is some indirect evidence. By studying transmission chains, in which participants pass information along to each other in sequence, Mesoudi, Whiten, and Dunbar (2006) demonstrated a bias toward transmitting and maintaining social information compared to

equivalent nonsocial information. Regarding reasoning, Mercier and Sperber (2010) argued that the ubiquity of reasoning errors supported their hypothesis regarding its argumentative functions. From this perspective, “errors” such as confirmation bias or examples of motivated reasoning (see Evans, Chapter 8; Molden & Higgins, Chapter 20) indicate that individuals do not reason in order to discover truth, but rather that reasoning is instead adapted to persuade listeners of our point of view. However, it should be noted that the mere existence of such biases does not constitute strong evidence for a social-functional view of reasoning, as they are also quite compatible with a Bayesian rational account of argumentation (see Hahn & Oaksford, Chapter 15; Rai & Holyoak, 2011).

Alternative social-functional theories of cognition with stronger support are Relational Models Theory (Fiske, 1991, 1992) and Relationship Regulation Theory (Rai & Fiske, 2011). According to Relational Models Theory, particular patterns of social interaction are stable across generations, and as a consequence we have evolved particular social-relational schemas for constituting and successfully navigating our social relationships. For example, “authority-ranking” relations are ubiquitous in societies and are typically constituted iconically through indices of magnitude, such as size or height, with greater magnitudes being associated with higher positions in the hierarchy. Whereas Relational Models Theory provides a taxonomy of the dominant social relationships that exist across cultures and how they emerge in ontogeny, Relationship Regulation Theory specifies the distinct moral motives and actions that have evolved to successfully maintain and regulate those relationships. For example, authority-ranking relations are motivated by “hierarchy,” which requires that subordinates respect and defer to superiors, who in turn are required to lead, guide, direct, and protect subordinates. (See Waldmann, Nagel, & Wiegmann, Chapter 19, for a review of work on moral judgment.)

Relational Models Theory predicts that people are predisposed to acquire certain social-relational schemas early in development and Relationship Regulation Theory predicts that people are predisposed to acquire corresponding moral motives. Thomsen et al. (2011) found that by 10 months of age, infants show greater surprise when they see a larger anthropomorphized block physically “bow” and defer to a smaller anthropomorphized block than vice versa, controlling for nonsocial factors.

As smaller and larger size are constitutive of subordinate and superior positions in authority-ranking relations, and as the hierarchy motive requires that subordinates defer to superiors, this finding suggests that even preverbal infants have expectations regarding appropriate social behavior in authority relationships, without having received explicit teaching. Similarly, infants as young as 8 months of age prefer a puppet that harmed a previously antisocial puppet to a puppet that helped that same puppet (Hamlin, Wynn, Bloom, & Mahajan, *in press*). This finding suggests that our minds are not innately predisposed toward specific moral actions, such as a prohibition against intentional harm (Haidt, 2007; Hauser, 2006); rather, we are predisposed to acquire expectations and motives regarding the correct conduct of social relationships without explicit teaching. Although “core knowledge” has been proposed in nonsocial domains such as intuitive physics or folk biology (Carey & Spelke, 1994), we are only now beginning to consider that the mind may also be predisposed toward early acquisition of social-relational knowledge necessary for successful living in sociocultural groups.

An Illusory Distinction

There is no such thing as a human nature independent of culture. Men without culture... would be unworkable monstrosities (Geertz, 1973, pp. 48–49).

Within classical formulations of psychology, a distinction has often been cast between what are considered conceptually important basic universal cognitive “processes,” such as perception, reasoning, decision making, and memory, and what are considered relatively unimportant learned “content” that informs those processes and varies across cultures. Thus, the Azande may believe in witchcraft, but it is not because they have engaged in cognitive processes fundamentally different from those of Westerners. Rather, their beliefs are due to what they have been taught; they find confirmation of their beliefs via the same reasoning biases that Westerners exhibit (Evans-Pritchard, 1937/1976). Conceptualized in terms of an analogy between the mind and an information-processing computer, cognitive processes represent underlying “computations,” whereas cultural content represents the “inputs” on which the computations are performed (Block, 1995). Different inputs will lead to different “outputs,” but the computation remains unchanged

and is therefore the fundamental object of study for investigators of cognition.

Nisbett and colleagues (reviewed in Nisbett, 2004; Nisbett et al., 2001) have challenged this view through their research into cross-cultural differences in cognition between individualistic (independent, analytic) and collectivist (interdependent, holistic) cultures. In line with contemporary theories of Bayesian rational reasoning and the importance of prior beliefs (see Griffiths, Tenenbaum, & Kemp, Chapter 3), they have contended that inputs into cognition are fundamental to cognitive processing because they determine what cognitive processes individuals will engage in. For example, one widely documented difference is between the “analytic” reasoning styles of individualist Westerners and the “holistic” reasoning styles of collectivist Easterners. Under this formulation, ecological and historical factors have led people in Western societies to assume direct cause-effect relations and consequently place greater emphasis on focal objects, whereas people in Eastern societies are more prone to assume that actions occur in a field of forces and consequently emphasize the wider context within which actions take place. In contrast to the Azande example, Nisbett et al. (2001) argue that due to these different implicit theories and epistemologies of the world, Easterners and Westerners will actually engage in different cognitive processes. For example, when viewing a scene, Easterners focus their attention on background images and have better memory for those images than Westerners, who in turn demonstrate the opposite pattern by focusing their attention on focal images in the foreground.

By Nisbett et al.’s (2001) own admission, their position is “at the same time less radical and more radical than the assertion that basic processes differ across cultures” (p. 306). On the one hand, they have suggested that culture does affect process in a profound way because the different theories and epistemologies learned in different cultures trigger different processes. On the other hand, they are still operating within the traditional conceptualization of a separation between content and process. They have simply argued for an additional computational step of selecting among multiple alternative cognitive processes in which an individual might engage. Cultural inputs that differ across cultures bias which cognitive processes are pursued. In this framework, culture is still operationalized as that which is different between two groups; it serves as a latent variable that can be used to explain additional variance, in

the same manner that social psychologists might use race or gender.

Although I agree that prior beliefs have been underemphasized and are certainly sensitive to culture, based on the research reviewed in this chapter I would like to push further by suggesting a blurrier (perhaps even nonexistent) distinction between process and content (Geertz, 1973). Living in large social groups connected by culture has selected for vast changes in neural systems and led to specialized psychology for succeeding in societies and cultures. Consequences include the evolution and development of cognitive capabilities specifically relevant to grasping implicit cultural meanings, including observational learning and imitation as well as biases that favor attending to social-relational information and thinking about social relationships. These cognitive capabilities have allowed us to build and inhabit more complex societies over time, enabling the distribution of knowledge across people, time, and technologies so that a given individual needs to know and think about incredibly little in order to function successfully.

In this framework, we cannot simply strip away culture to see what cognitive capacities are left. We will find relatively little remaining, as much of the human mind has adapted specifically for functioning in societies and cultures. Rather, we must investigate cognition in the context of the modes of thought produced by evolving and living in societies and cultures.

Conclusions and Future Directions

My aim in this chapter has been to raise some basic questions about how thinking is manifested within societies and cultures. In so doing, I have suggested that we should shift our attention toward examining cognitive processes at the level of the social group and more fully integrate analyses of the sociocultural environment into psychological theorizing. The promise of this approach is that it may provide insights into cognition that could never be attained if we focused only on the cognitive processes that lend themselves to a view of individual minds working in isolation.

In modern economics there is a distinction between microlevels and macrolevels of analysis that captures the idea that patterns and causal relations operating at the level of large groups and societies cannot simply be reduced to the aggregate principles underlying the actions of individuals. Rather, there are emergent properties of groups that must

be analyzed using different methods and theories. Within psychology, research on cognition has been dominated by investigations into the processes that occur within an individual mind studied in isolation. The research discussed in this chapter can be interpreted as a proposal for establishing a division between such micropsychology and the macropsychology discussed in this chapter. This division is based on the recognition that individual minds are part of larger cognitive networks, which have emergent psychological properties that cannot be reduced to individual psychological processes alone. In this framework, we will have to consider how social ecologies impact patterns of social relations, generating cultural histories that in turn produce particular patterns of cognition at both the microlevel of individuals and the macrolevel of groups.

This new model of research will require that we be willing to leave the laboratory and inductively develop our theories based on real-world observation of everyday social interaction (Rai & Fiske, 2010). For example, at the micropsychological level, cognitive dissonance (Festinger & Carlsmith, 1959), which was interpreted in terms of the anxiety felt by individuals who act contrary to their beliefs, proposed that people will change their beliefs to align with their actions in order to reduce anxiety. For such a process to occur, individuals presumably must believe that actions are freely chosen, that their choices imply preferences that are stable across time and reflect their character, that they have control over and are thus responsible for the outcome of events, and that good people make consistent choices. However, none of these assumptions are cross-culturally universal (Fiske et al., 1998). Not coincidentally, Japanese participants do not display dissonance reduction in the free-choice paradigm (Heine & Lehman, 1997). By integrating an understanding of the individual's psychological embeddedness within a sociocultural context, we will be able to generate stronger causal theories of behavior as well as use cultural differences as a test bed for assessing these theories.

At the macropsychological level, areas of cognitive investigation that have heretofore been concentrated in disciplines outside of psychology, such as analyses of collective memory in sociology, cultural transmission in evolutionary anthropology, social-collaborative technologies in computer science, distributed cognition of complex tasks, and the consequences of biological changes in response to the sociocultural environment, become principal objects of

psychological inquiry. Particular questions to address will focus on how knowledge is represented at the group level. For example, many scientists wish to discover what memory is “in the brain.” But what is memory “in the group”? How will we describe it? In this chapter I have suggested that technologies such as computers or books can be thought of as external memory storage. But what about aspects of our environment, such as landmarks or a worn path in the wilderness? I may have no deep understanding of how to get to where I wish to go, but I know if I follow the path I will get there. Should we consider the knowledge to be encoded into the path, or is the path simply a heuristic that my individual mind exploits?

Of course, it is unlikely that all knowledge is distributed. Indeed, there likely are instances in which no one actually knows (or has ever known, even in an unconscious fashion), and yet thinking occurs without knowledge. For example, some practices and traditions may have initially occurred at random, but then were maintained and transmitted over time because they served some function of which their progenitors were unaware. One question to consider is how much of our cognition reflects such thinking without knowledge.

Finally, single nucleotide polymorphisms in our genes are highly susceptible to local changes in the environment (see Green & Dunbar, Chapter 7). Their incidence has increased drastically in humans over the last 10,000 years, most likely due to rapid population growth that occurred as a consequence of the cultural evolution of large social group living (Cochran & Harpending, 2010). Although in its early stages and quite speculative, recent work suggests that differences in gene expression as a result of polymorphisms are linked to differences in psychological functioning across cultures, raising the question of whether differences in gene expression may be occurring as a function of cultural change (Kitayama & Uskul, 2011).

Acknowledgments

Preparation of this chapter was supported by the UCLA Center for Society and Genetics. I thank Clark Barrett, Patricia Cheng, Daniel Fessler, Alan Fiske, Keith Holyoak, the UCLA Reasoning Lab, the UCLA Relational Models Theory lab, and the UCLA Experimental Biological Anthropology Lab for their valuable comments on an earlier draft.

Notes

1. Note that my use of *macropsychology* and *micropsychology* is quite different from that of Fiske (1991) and from conceptions directly tied to economics (Katona, 1979).

2. To rule out the “arrogant Yale graduate student” hypothesis, the results were replicated with non-Ivy league undergraduates.

3. Although not the focus of this chapter, both individual and social learning are necessary for cumulative cultural evolution to evolve.

4. Wikipedia is an interesting example. Although it may have originated in a very free-flowing manner, it now has layers of higher level editing. Similar hierarchical structures may also emerge in other forms of social collaboration.

5. Studies that reveal “activation” in these areas in response to general cognitive tasks are actually reporting *less deactivation* from the resting baseline compared to control tasks, such as staring at a fixation point.

References

- Anderson, L., & Holt, C. (1997). Information cascades in the laboratory. *The American Economic Review*, 87, 847–862.
- Anderson, M. (2003). Embodied cognition: A field guide. *Artificial Intelligence*, 149, 91–130.
- Asch, S. (1955). Opinions and social pressure. *Scientific American*, 193, 31–35.
- Austin, J. (2003). Transactive memory in organizational groups: The effects of content, consensus, specialization, and accuracy on group performance. *Journal of Applied Psychology*, 88, 866–878.
- Bandura, A., Ross, D., & Ross, S. (1963). Imitation of film-mediated aggressive models. *Journal of Abnormal and Social Psychology*, 66, 3–11.
- Bertrand, M., Mullainathan, S., & Shafir, E. (2004). A behavioral-economics view of poverty. *The American Economic Review*, 94, 419–423.
- Block, N. (1995). The mind as the software of the brain. In E. Smith & D. Osherson (Eds.), *Thinking: An invitation to the cognitive sciences* (pp. 377–425). Cambridge, MA: MIT Press.
- Boyer, P. (1992). *Tradition as truth and communication*. Cambridge, England: Cambridge University Press.
- Bray, R., & Noble, A. (1978). Authoritarianism and decisions of mock juries: evidence of jury bias and group polarization. *Journal of Personality and Social Psychology*, 36, 1424–1430.
- Brooks, D. (2007, October, 26). The outsourced brain. *The New York Times*, Retrieved September 2011, from <http://www.nytimes.com/2007/10/26/opinion/26brooks.html>
- Brown, J., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, 18, 32–42.
- Caldwell, C., & Millen, A. (2008). Experimental models for testing hypotheses about cumulative cultural evolution. *Evolution and Human Behavior*, 29, 165–171.
- Carey, S., & Spelke, E. (1994). Domain-specific knowledge and conceptual change. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain Specificity in Culture and Cognition* (pp. 169–200). Cambridge: Cambridge University Press.
- Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58, 7–19.
- Cochran, G., & Harpending, H. (2010). *The 10,000 year explosion: How civilization accelerated human evolution*. New York, NY: Basic Books.
- Cole, M. (1995). Culture and cognitive development: From cross-cultural research to creating systems of cultural mediation. *Culture and Psychology*, 1, 25–54.

- D'Andrade, R. G. (1981). The cultural part of cognition. *Cognitive Science*, 5, 179–195.
- Dunbar, R. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22, 469–493.
- Dunbar, R. (1993). Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences*, 16, 681–735.
- Dunbar, R. (1996). *Grooming, gossip, and the evolution of language*. Cambridge, MA: Harvard University Press.
- Dunbar, R., & Shultz, S. (2007). Evolution in the social brain. *Science*, 317, 1344–1347.
- Evans-Pritchard, E. (1976). *Witchcraft, oracles, and magic among the Azande*. Oxford, England: Clarendon University Press. (Original work published 1937).
- Fessler, D., & Machery, E. (2012). Culture and cognition. In E. Margolis, R. Samuels, & S. Stich (Eds.), *The Oxford handbook of philosophy of cognitive science* (pp. 503–527). Oxford, England: Oxford University Press.
- Festinger, L., & Carlsmith, J. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology*, 58, 203–210.
- Fiske, A. (1991). *Structures of social life*. New York: Free Press.
- Fiske, A. (1992). Four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review*, 99, 689–723.
- Fiske, A., Kitayama, S., Markus, H., & Nisbett, R. (1998). The cultural matrix of social psychology. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., pp. 414–481). New York: McGraw-Hill.
- Fiske, D. (1991). Macropsychology and micropsychology: Natural categories and natural kinds. In R. Snow & D. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 61–74). Hillsdale, NJ: Erlbaum.
- Fiske, S. (1992). Thinking is for doing: Portraits of social cognition from Daguerreotype to laserphoto. *Journal of Personality and Social Psychology*, 63, 877–889.
- Gallese V., & Goldman A., (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, 2, 493–501.
- Geertz, C. (1973). *The interpretation of cultures: Selected essays*. New York: Basic Books.
- Gergely, G., Bekkering, H., & Kiraly, I. (2002). Rational imitation in preverbal infants. *Nature*, 415, 755.
- Griffiths, T., Kalish, M., & Lewandowsky, S. (2008). Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philosophical Transactions of the Royal Society B*, 363, 3503–3514.
- Haidt, J. (2007). The new synthesis in moral psychology. *Science*, 316, 998–1002.
- Halbwachs, M. (1950). *On collective memory*. New York: Harper Colophon Books.
- Hamlin, J.K., Wynn, K., Bloom, P., & Mahajan, N. (in press). The richness of early social evaluation. *Proceedings of the National Academy of Science*.
- Harris, J. (1995). Where is the child's environment? A group socialization theory of child development. *Psychological Review*, 102, 458–489.
- Hauser, M. (2006). *Moral minds: How nature designed our universal sense of right and wrong*. New York: Ecco Books.
- Heine, S., & Lehman, D. (1997). Culture, dissonance, and self-affirmation. *Personality and Social Psychology Bulletin*, 23, 389–400.
- Heine, S., & Ruby, M. (2010). Cultural psychology. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2, 254–266.
- Henrich, J. (2004). Demography and cultural evolution: How adaptive cultural processes can produce maladaptive losses: The Tasmanian case. *American Antiquity*, 69, 197–214.
- Henrich, J., & Boyd, R. (1998). The evolution of conformist transmission and the emergence of between-group differences. *Evolution and Human Behavior*, 19, 215–241.
- Henrich, J., & Gil-White, F. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior*, 22, 165–196.
- Henrich, J., & McElreath, R. (2003). The evolution of cultural evolution. *Evolutionary Anthropology*, 12, 123–135.
- Hollingshead, A. (1998). Communication, learning, and retrieval in transactive memory systems. *Journal of Experimental Social Psychology*, 34, 423–442.
- Hutchins, E. (1995a). *Cognition in the wild*. Cambridge, England: Cambridge University Press.
- Hutchins, E. (1995b). How a cockpit remembers its speeds. *Cognitive Science*, 19, 265–288.
- Iacoboni, M., Lieberman, M., Knowlton, B., Molnar-Szakacs, I., Moritz, M., Throop, J., & Fiske, A. (2004). Watching social interactions produces dorsomedial prefrontal and medial parietal BOLD fMRI signal increases compared to a resting baseline. *Neuroimage*, 21, 1167–1173.
- Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J., & Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology*, 3, 529–535.
- Isenberg, J. (1986). Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology*, 50, 1141–1151.
- Katona, G. (1979). Toward a macropsychology. *American Psychologist*, 34, 118–126.
- Keil, F. (2003). Folkscience: Course interpretations of a complex reality. *Trends in Cognitive Science*, 7, 368–373.
- Keil, F. (2006). Explanation and understanding. *Annual Review of Psychology*, 57, 227–254.
- Keil, F. (2010). The feasibility of folk science. *Cognitive Science*, 34, 826–862.
- Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, 18, 513–549.
- Kitayama, S., & Uskul, A. (2011). Culture, mind, and the brain: Current evidence and future directions. *Annual Review of Psychology*, 62, 419–449.
- Kittur, A., & Kraut, R. (2008). Harnessing the wisdom of crowds in wikipedia: Quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work* (pp. 37–46). New York: ACM Press.
- Lancy, D. (1996). *Playing on the mother-ground: Cultural routines for children's development*. New York: Guilford Press.
- Latane, B. (1981). The psychology of social impact. *American Psychologist*, 36, 343–356.
- Latane, B. (1996). Dynamic social impact: The creation of culture by communication. *Journal of Communication*, 46, 13–25.
- Latane, B., & Bourgeois, M. (1996). Experimental evidence for dynamic social impact: The emergence of subcultures in electronic groups. *Journal of Communication*, 46, 35–47.
- Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, England: Cambridge University Press.

- Lehman, D., Chiu, C., & Schaller, M. (2004). Psychology and culture. *Annual Review of Psychology*, 55, 689–714.
- Liang, D., Moreland, R., & Argote, L. (1995). Group versus individual training and group performance: The mediating role of transactive memory. *Personality and Social Psychology Bulletin*, 21, 384–393.
- Lillard, A. (1998). Ethnopsychologies: Cultural variations in theory of mind. *Psychological Bulletin*, 123, 3–32.
- Loftus, E., & Hoffman, H. (1989). Misinformation and memory: The creation of new memories. *Journal of Experimental Psychology: General*, 118, 100–104.
- Lughod, L. (1991). Writing against culture. In R. Fox (Ed.), *Recapturing anthropology: Working in the present* (pp. 137–154). Santa Fe, NM: School of American Research.
- Markus, H., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, 98, 224–253.
- Meltzoff, A., & Moore, M. (1977). Imitation of facial and manual gestures by human neonates. *Science*, 198, 75–78.
- Mercier, H., & Sperber, D. (2010). Why do humans reason? Arguments for an argumentative theory. *Brain and Behavioral Sciences*, 34, 57–74.
- Mesoudi, A., Whiten, A., & Dunbar, R. (2006). A bias for social information in human cultural transmission. *British Journal of Psychology*, 97, 405–423.
- Miller, D., & Prentice, D. (1994). Collective errors and errors about the collective. *Personality and Social Psychology Bulletin*, 20, 541–550.
- Moscovici, S., & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of Personality and Social Psychology*, 12, 125–135.
- Nisbett, R. (2004). *The geography of thought: How Asians and Westerners think differently and why*. New York: Free Press.
- Nisbett, R., & Norenzayan, A. (2002). Culture and cognition. In D. Medin (Ed.), *Stevens's handbook of experimental psychology: Cognition*. (3rd ed., p. 561–597). New York: Wiley.
- Nisbett, R., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review*, 108, 291–310.
- Olick, J., & Robbins, J. (1998). Social memory studies: From “collective memory” to the historical sociology of mnemonic practices. *Annual Review of Sociology*, 24, 105–140.
- Onishi, K., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308, 255–258.
- Rai, T., & Fiske, A. (2010). Psychological studies are ODD (observation and description deprived). Commentary in *Brain and Behavioral Sciences*, 33, 106–107.
- Rai, T., & Fiske, A. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review*, 118, 57–75.
- Rai, T., & Holyoak, K. (2011). The rational hypocrite: A Bayesian approach to moral hypocrisy. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*.
- Reader, S., & Laland, K. (2002). Social intelligence, innovation, and enhanced brain size in primates. *Proceedings of the National Academy of Sciences USA*, 99, 4436–4441.
- Richerson, P., & Boyd, R. (2005). *Not by genes alone: How culture transformed human evolution*. Chicago, IL: University of Chicago Press.
- Richerson, P., Boyd, R., & Henrich, J. (2010). Gene-culture co-evolution in the age of genomics. *Proceedings of the National Academy of Sciences USA*, 107, 8985–8992.
- Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27, 169–192.
- Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, 26, 521–562.
- Schwartz, B. (1991). Social change and collective memory: The democratization of George Washington. *American Sociological Review*, 56, 221–236.
- Segall, M., Lonner, W., & Berry, J. (1998). *American Psychologist*, 53, 1101–1110.
- Sperber, D. (1985). Anthropology and psychology: Toward an epidemiology of representations. *Man*, 20, 73–89.
- Steele, C., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African-Americans. *Journal of Personality and Social Psychology*, 69, 797–811.
- Suwiecki, J. (2005). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes businesses, economies, societies, and nations*. New York: Doubleday.
- Thomsen, L., Frankenhuys, W., Ingold-Smith, M., & Carey, S. (2011). Big and mighty: Preverbal infants mentally represent social dominance. *Science*, 331, 477–480.
- Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.
- Triandis, H. (1989). The self and social behavior in differing cultural contexts. *Psychological Review*, 96, 506–520.
- Twain, M. (1889). *A Connecticut Yankee in King Arthur's court*. New York: Harper.
- Vygotsky, L. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.
- Wegner, D. (1987). Transactive memory: A contemporary analysis of the group mind. In B. Mullen & G. Goethals (Eds.), *Theories of group behavior* (pp. 185–208). New York: Springer-Verlag.
- Weldon, M., & Bellinger, K. (1997). Collective memory: Collaborative and individual processes in remembering. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 1160–1175.
- Welzer, H. (2010). Re-narrations: How pasts change in conversational remembering. *Memory studies*, 3, 5–17.
- Wilson, R. (2005). Collective memory, group minds, and the extended mind thesis. *Cognitive Processing*, 6, 227–236.
- Wilson, R., & Keil, F. (1998). The shadows and shallows of explanation. *Minds and Machines*, 8, 137–159.
- Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives*, 18, 107–126.
- Zerubavel, E. (1996). Social memories: Steps to a sociology of the past. *Qualitative Sociology*, 19, 283–299.

PART

6

Modes of Thinking

This page intentionally left blank

Development of Quantitative Thinking

John E. Opfer and Robert S. Siegler

Abstract

For understanding development of quantitative thinking, the distinction between nonsymbolic and symbolic thinking is fundamental. Nonsymbolic quantitative thinking is present in early infancy, culturally universal, and similar across species. These similarities include the ability to represent and compare numerosities, the representations being noisy and increasing logarithmically with actual quantity, and the neural correlates of number representation being distributed in homologous regions of frontoparietal cortex. Symbolic quantitative thinking, in contrast, emerged recently in human history, differs dramatically across cultural groups, and develops over many years. As young children gain experience with symbols in a given numeric range and associate them with nonverbal quantities in that range, they initially map them to a logarithmically-compressed mental number line and later to a linear form. This logarithmic-to-linear shift expands children's quantitative skills profoundly, including their ability to estimate positions of numbers on number lines, to estimate measurements of continuous and discrete quantities, to categorize numbers by size, to remember numbers, and to estimate and learn answers to arithmetic problems. Thus, while nonsymbolic quantitative thinking is important and foundational for symbolic numerical capabilities, the capacity to represent symbolic quantities offers crucial cognitive advantages.

Key Words: numerical cognition, number representation, mathematical thinking, symbols, cognitive development

Quantitative thinking is central to human life. Whether the situation involves a child recalling which blocks provided her with the most candies on previous Halloweens, a Londoner telling the time by counting the tolls of Big Ben, or a candidate using polls to predict results of an upcoming election, quantitative thinking is important for learning from the past, monitoring the present, and planning for the future.

Quantitative thinking plays an important role in the lives of other animals as well. To project the outcome of a future fight, prides of lionesses compare their pridenumbers to the number of distinct roars they hear in rival packs (McComb, Packer, & Pusey, 1994). Similarly, to learn optimal foraging

locations, animals in the wild encode the relative number of food items they have found previously in various locations (Davis, 1993).

The existence of such quantitative abilities makes sense from an evolutionary perspective. Being deprived of a sense of *how many* would deprive an animal of any rationality in its judgments and decisionmaking. Rational choices among alternative strategies and courses of action would be rendered impossible.

Given the importance of quantitative thinking, it is unsurprising that representations of quantity are a universal property of human cognition. Quantitative representations are present from early infancy (Cordes & Brannon, 2008a; Feigenson, Dehaene, &

Spelke, 2004) and share striking similarities across human cultures that provide radically different cultural and linguistic experiences (Butterworth et al., 2008; Gordon, 2004; Pica, Lemer, Izard, & Dehaene, 2004). Moreover, early-developing quantitative representations play a central role in learning a wide range of other quantitative skills that have emerged more recently in human history, such as symbolic arithmetic and algebra (Ifrah, 2000).

The central distinction among numerical capabilities that organizes this review, one that is highly correlated with the distinction between early-developing and later-developing quantitative skills, is that between nonsymbolic and symbolic capabilities. Nonsymbolic capabilities are ones in which numerical properties are implicit; for example, when sets of 7 and 5 dots are presented, the facts that one set has 7, the other has five, and the first set has more objects than the second are all implicit. In contrast, symbolic capabilities are ones in which the numerical properties are explicitly expressed as written Arabic numerals, spoken words, or written words (e.g., “7” or “seven”). Nonsymbolic and symbolic numerical processing differ in many ways. Nonsymbolic processing of numbers is widespread across species; symbolic numerical processing appears to be unique to humans (aside from a small number of primates that have participated in laboratory experiments). Nonsymbolic processes emerge in infancy; symbolic processing does not emerge until later in childhood. Nonsymbolic processing is approximate; symbolic processing allows precision. Moreover, symbolic processing of numbers shows wide variations among different cultures and historical periods, whereas variation in nonsymbolic processing is less marked.

One reason to be interested in the development of quantitative abilities is that findings from this area often shed light on general theoretical issues in cognitive development. These issues include the potential existence of innate representational abilities, the extent to which early-developed capacities are sufficient to support the development of later abilities, and whether experience creates new representational resources or selects among preexisting ones for use in novel contexts. Among the reasons why studies of quantitative abilities have been so fruitful is that such studies permit a degree of mathematical precision that is much more typical of psychophysics than of studies of children’s concepts in other areas (e.g., theory of mind, biology, and moral development), thereby allowing researchers to test models

that generate competing quantitative predictions. An equally important reason for the rapid expansion of this area is that valuable practical applications for improving children’s mathematical understanding have arisen from the theoretical work.

Within the distinction between nonsymbolic and symbolic quantitative thinking, we focus on three main questions: (1) What quantitative abilities exist in infancy that provide the foundation for later, more advanced abilities; (2) How are later abilities continuous, and how are they discontinuous, with these early emerging abilities; and (3) How do developmental changes in these abilities affect other aspects of quantitative thinking, such as mental arithmetic?

Nonsymbolic Quantitative Thinking

Foundations of Quantitative Thinking in Infancy

Three key issues arise in studying the early foundations of human nonsymbolic quantitative abilities. First, at what point in development do children reliably *discriminate* among quantities that differ in number and *represent* the number of entities in a set? Second, in early development, what mental mechanisms initially represent numerical values? Third, at what point in development can children *mentally manipulate* numbers (as measured, for example, by their ability to recognize impossible arithmetic transformations)? It turns out that all of these abilities are present early in infancy.

DISCRIMINATION OF NUMERICAL QUANTITIES

Without the ability to perceive the difference between two sets that differ only in number, it would be impossible to understand the similarity of sets that have only number in common (cross-modal mapping), to distinguish between possible and impossible arithmetic transformations of sets of objects (nonsymbolic arithmetic), to link numeric symbols to their approximate or exact referents (estimation, counting), or to engage in economic transactions. For these reasons, research on the development of numerical cognition begins with efforts to establish the capacities (and limits) of infants’ discrimination among quantities.

Infants’ Discrimination Among Numbers of Objects

A consistent finding in research on development of quantitative abilities is that human infants notice changes in number in sets of 1–3 objects (Starkey & Cooper, 1980; Starkey, Spelke, & Gelman, 1983; Strauss & Curtis, 1981). In one early study of

infants' numerical capacities and a number of later ones, researchers observed a spontaneous preference for the larger set when two sets differed in number by a factor of 2 or more, even with very large sets (e.g., 128 versus 32 elements in Fantz & Fagan, 1975; 32 versus 16 in Xu, Spelke, & Goddard, 2005; 16 versus 8 in Lipton & Spelke, 2003). In other studies, investigators used habituation paradigms in which infants were repeatedly presented a particular number of objects until their looking time decreased, and then they were presented with a different number of objects. Recovery of looking time to novel stimuli (dishabituation) was observed when the ratio between original and novel sets was 2:1 or 3:2 (2 versus 3 in Strauss & Curtis, 1981; 8 versus 4 in Xu, 2003). The earliest signs of this kind of numerical discrimination appeared in 21- to 44-hour-old neonates, who—having been habituated to a display of 2 (or 3) dots—recovered interest when shown 3 (or 2) dots (Antell & Keating, 1983).

Findings from studies of infants' discrimination of sets of objects suggest three general conclusions. First, at all ages, the number required to discriminate between two relatively large sets (i.e., 4 or more members) is not an absolute number but rather a ratio (e.g., discrimination of 4 versus 8 objects is typical at ages where discrimination of 8 versus 12 objects is not). Rather, as in the Weber-Fechner psychophysical function, *the probability of discrimination is proportional to the difference in the logarithms of the numbers*, where, for example, $\ln(32)-\ln(16) = \ln(16)-\ln(8) = \ln(8)-\ln(4) > \ln(12)-\ln(8) = \ln(6)-\ln(4)$. Second, the difference in logarithms required to discriminate between numbers decreases with age. Thus, older infants discriminate ratios that younger infants do not (Cordes & Brannon, 2008b). Finally, for very small numbers (i.e., 3 or less), the probability of discrimination is uniformly high, higher than would be predicted from considering the differences in logarithms alone. Thus, discriminating 2 versus 3 is easier for infants than discriminating 4 versus 6.

Infants' discrimination of auditory sequences (Lipton & Spelke, 2003; vanMarle & Wynn, 2009), temporal intervals (Brannon et al., 2008; vanMarle & Wynn, 2006), events (Wynn, 1996; Wood & Spelke, 2005,) and collective entities (Wynn et al., 2002) conforms to these generalizations about number discrimination. To cite one example, the Weber-Fechner law applies to discriminations among number of sounds. When Lipton and Spelke (2003) familiarized 6-month-old infants to a sequence of 8 or 16 sounds,

infants more often turned their heads to hear a novel number of sounds (16 or 8) than to hear the original number of sounds. As with objects, the ratio of the sets is what matters; 6-month-olds discriminate between 32 and 16, 16 and 8, and 8 and 4 sounds, but not between 4 and 6 or 8 and 12 sounds (Lipton & Spelke, 2004; Xu, 2003; Xu et al., 2005). Also as with objects, the difference in logarithms required to discriminate between two numbers of sounds decreased with age: Older infants discriminated ratios (e.g., 4:6) that younger infants did not (Lipton & Spelke, 2003). And as with objects, discrimination of very small numbers of sounds is consistently high, higher than would be expected from considering the ratio in isolation. When 4-day-old infants were presented multisyllabic utterances, discriminations between 2 and 3 syllables were more likely than between 4 and 6 (Bijeljac-Babic, Bertoni, & Mehler, 1993).

DO INFANTS REPRESENT NUMBER PER SE?

Whenever infants react to changes in a stimulus, a number of possible mechanisms might give rise to the reaction. In the case of numerical discrimination, different mechanisms might process numbers of discrete entities than other quantitative dimensions that often are correlated with number of discrete entities, such as summed area of the objects or their contour length. Because numerical and non-numerical parameters are inextricably linked in sets of objects, it is sometimes impossible to uniquely identify the factors that allowed infants to dishabituate to a display within a specific experiment (Mix, Huttenlocher, & Levine, 2002). Thus, infants' ability to discriminate sets of items that differ in number may not necessarily mean that infants possess mental mechanisms that code number per se. In the next two sections, we examine this issue, first by reviewing whether infants' sensitivity to nonnumerical features of sets is sufficient to explain their ability to discriminate between sets that differ numerically, and then by examining evidence from cross-modal mapping studies that we view as decisive on this issue.

Nonnumerical Cues to Numerical Quantity

One class of nonnumerical cues that might cause an infant to dishabituate to a novel number of objects is continuous quantitative cues, such as surface area, volume, and contour length. A number of investigators have posited that discrimination among values of these continuous dimensions, rather than of number as such, explains findings that others have interpreted as indicative of numerical

discrimination abilities (Clearfield & Mix, 1999, 2001; Feigenson, Spelke, & Carey, 2002; Gao, Levine, & Huttenlocher, 2000; Mix, Huttenlocher, & Levine, 2002). In an important challenge to early research on infants' numerical abilities, Clearfield and Mix (1999) habituated 6-month-olds to a series of stimuli that shared a set size (e.g., three objects) and a constant cumulative contour length. Infants were then presented a dishabituation trial, either with number held constant and cumulative contour length changed or vice versa. Infants responded as if they detected the change in contour length but not the change in number. Results of subsequent studies indicated that infants' discrimination of surface area resembles that of the Weber-Fechner psychophysical function, with novelty preference increasing linearly with the ratio of the surface areas, a result that paralleled previous findings with discrete objects (Clearfield & Mix, 2001). One interpretation of these results (e.g., Mix et al., 2002; Newcombe, 2002) is that young infants do not represent number and instead respond solely to nonnumerical properties of the set.

On the other hand, several considerations suggest that nonnumerical cues are insufficient to account for infants' ability to discriminate sets of objects. First, when nonnumerical properties of a large set of objects (e.g., eight or more) are varied in the habituation display while number is held constant, infants in the test phase of the procedure reliably dishabituate when shown a set with a novel number of objects that shares nonnumerical properties of the habituation sets (Brannon, Abbott, & Lutz, 2004; Brannon, Lutz, & Cordes, 2006; Xu & Spelke, 2000). If infants were sensitive only to the nonnumerical properties of a set, this pattern of results would not occur. Second, discriminating nonnumerical values of a set is often more difficult for infants than discriminating the numbers of objects in the set (Brannon et al., 2004; Cordes & Brannon, 2008b, 2009). In Brannon et al.'s (2004) study, for example, 6-month-olds detected a two-fold change in number after being habituated to a five-fold change in surface area of the objects, but failed to detect a two-fold change in surface area of the objects after being habituated to a five-fold change in their number. Third, the similarity of findings on visual and auditory stimuli cannot be explained by spatial dimensions, such as contour length and surface area, which do not apply to sounds. Thus, rather than number being detected as a last resort, it seems more likely that infants

simultaneously track the number of discrete objects in a set and the nonnumerical characteristics of those objects. Consistent with this interpretation, combining auditory and visual cues to number improves the Weber-Fechner ratio that infants can discriminate over that which they can discriminate on the basis of visual cues alone (Jordan, Suanda, & Brannon, 2008).

Numerical Representation in Infants: Evidence From Cross-Modal Mapping

When infants are habituated to a display of eight objects or eight tones and then recover interest in a display of four objects or four tones, is it because they recognize that the "eighthness" of the habituation displays had been violated by the "fourness" of the test displays? This question is crucial because it gets at the heart of whether infants have a concept like "four" or "eight," as when adults generalize the word "four" to four votes, four belltolls, and four candies. To determine whether infants also have an abstract concept to which adults and older children would sensibly apply number words, researchers have tested infants' ability to generalize across numerical groups that—like candies and bell-tolls—are perceived by different sensory modalities (Izard, Sann, Spelke, & Streri, 2009; Jordan & Brannon, 2006; Kobayashi, Hiraki, & Hasegawa, 2005; Lourenco & Longo, 2010). An important feature of these cross-modal mapping tests is that they obviate objections that infants' reactions are driven only by nonnumerical features of the input—contour length usually increases with the number of candies, but it doesn't increase with the number of bell tolls. Similarly, temporal length of a sequence of bell tolls generally increases with number of tolls but not with objects' contour length.

In cross-modal mapping studies, infants were initially reported to look longer at a set of 2–3 objects that matched the number of sounds that were played simultaneously (Starkey, Spelke, & Gelman, 1983). However, subsequent studies reported no such preference (Mix et al., 1997) or a reversed preference (Mix et al., 1997; Moore et al., 1987). In all cases, the effect sizes were quite small, leading to the hypothesis that the contradictory findings were due to participants on the tails of a distribution whose means were in the middle (Mix et al., 2003).

Recent studies, however, have reported robust evidence of cross-modal number matching for sets of 2–3 s in 6-month-olds (objects and tones in Kobayashi et al., 2005; faces and voices in Jordan & Brannon, 2006) and of larger sets in

2-day-olds (Izard et al., 2009). An interesting feature of Izard et al. (2009) is that newborns' looking time was greater when viewing congruent displays (4 syllables/4 objects or 12 syllables/12 objects) than incongruent displays (4 syllables/12 objects or 12 syllables/4 objects). The difference in that study between congruent and incongruent displays was nearly identical for another three-fold change in number (6 vs. 18), and both differences were much greater than those evoked by a two-fold change in number (4 vs. 8). These findings are thought provoking, because they suggest that when newborns represent number in different modalities, their number representations are subject to the same ratio dependence observed in number discrimination within a single modality.

Origins of the Mental Number Line

If infants are capable of representing the number of items in a set, how might they do it? An important proposal for how this occurs—the mental number-line hypothesis—came from findings on number matching in nonhuman animals (Mechner, 1958; Platt & Johnson, 1971; for review, see Boysen & Capaldi, 1993).

PARALLELS BETWEEN INFANTS' AND NONHUMAN ANIMALS' NUMERICAL MAGNITUDE PROCESSING

To examine rats' matching of number of physical actions to a criterion number (i.e., the number of actions rewarded on previous trials), Mechner (1958) and Platt and Johnson (1971) developed paradigms in which food was dispensed to a rat after it had pressed a lever a criterion number of times and then stopped pressing the lever, either in order to press another lever (Mechner, 1958) or to enter a feeding area (Platt & Johnson, 1971). Results of the two studies were similar, with the modal number (and standard deviation) of lever presses increasing with the criterion number (4, 8, 12, and 16). To ensure that the rats were estimating the number of lever presses, rather than the time since the trial had started, subsequent studies varied the degree of food deprivation imposed on the rats, on the logic that the hungriest rats would press the lever faster than the less-hungry rats (which they did). Speed of bar pressing did not affect matching of bar presses to the criterion number; the hungry rats still pressed the lever the same *number* of times as the less-hungry rats for each criterion number (Mechner & Guevrekian, 1962).

The similarity of rats' numerical processing to that of infants can be seen in the rats' pattern of errors. Specifically, the likelihood of pressing the lever a given number of times in response to a particular criterion number decreased *as a function of the difference in the two numbers' logarithms*. For example, as in the previously described studies of human infants, Mechner found that rats were more likely to confuse 6 with 4 lever presses than to confuse 8 with 4, and rats were more likely to confuse 16 with 12 than 8 with 4. As Gallistel (1993) wrote, "It is as if the rat represented numerosity as a position on a [logarithmic!] mental number line (that is, a continuous mental/neural quantity or magnitude), using a noisy mapping process from numerosity to values on this continuum, so that the one and same numerosity would be represented by somewhat different mental magnitudes (positions on the number line) on separate occasions" (p. 215).

Why might a "mental number line" (illustrated in Fig. 30.1) be useful for thinking about infants' and nonhuman animals' encoding of number? One reason is that the mental number line provides a nonverbal mechanism for a true number concept, thereby accounting for prelinguistic infants' (and other animals') ability to generalize across perceptual modalities (Dehaene, 1992). If infants automatically convert the sight of six apples or the sound of six tones (Fig. 30.1, bottom panel, "Stimuli") to the same position on a mental number line (Fig. 30.1, middle panel, "Mental Number Line"), then—long before they learn conventional verbal counting—they could recognize an equivalent "sixness" of the two sets, to which a written or oral symbol ("six" or "6") could later be associated.

The mental number line metaphor is also useful for illustrating how number discrimination follows the Weber-Fechner law. That is, if mental magnitudes increase logarithmically with actual numeric value (as numbers on a slide rule), then numeric intervals on the lower end of the number line (like 1 and 2) are further apart and—given noisiness in the mapping—easier to discriminate than the same intervals at higher ends of the range, due to reduced overlap of signals.

The idea that the mapping of quantities to the number line is noisy provides an interesting way to think about what changes in infants' development. If the noisiness of the mappings to the logarithmic number line decreases with age, then the difference in logarithms required to discriminate any two quantities would also decrease with age. Thus, a

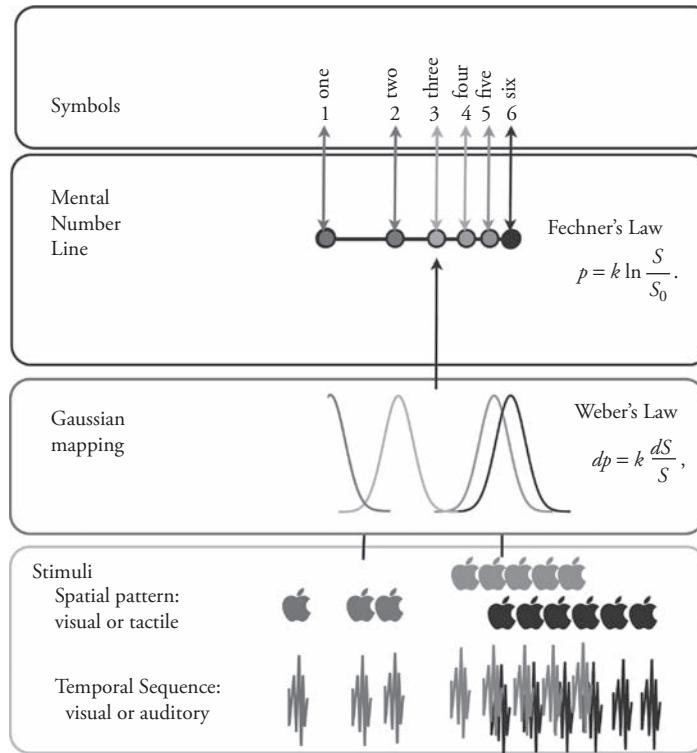


Fig. 30.1 Illustration of a logarithmically-scaled mental number line. See color figure.

logarithmically-scaled mental number line provides a simple way to conceptualize the numerical abilities of infants, how those abilities are limited by the Weber-Fechner Law, and how those abilities change with age and experience.

THE NEURAL BASIS OF THE MENTAL NUMBER LINE

What physical mechanisms generate the findings that are explained metaphorically by the mental number line construct? An idea that proved prescient was formulated by Dehaene and Changeux (1993). Working from the pattern of errors and solution times in numerosity discrimination studies that we reviewed earlier, as well as Dehaene's own research on symbolic number comparison in human adults, Dehaene and Changeux proposed the existence of innate numerosity detectors, that is, neural units that directly code numerosity (cf. "numerons" in Gelman & Gallistel, 1978).

In the Dehaene-Changeux model, objects of various sizes and locations are initially presented to a retina-like module (or to echoic auditory memory in the case of sounds), and then normalized for size and location on an intermediate topographical map of object locations. Number is registered from a

map of numerosity detectors that sum all outputs from the intermediate topographical map of object locations. An important feature of such a mechanism is that the internal representation of numerosity would be highly correlated with the number of objects in a set, regardless of their physical characteristics. Additionally, simulations that embodied the model revealed that the activations evoked by different input numerosities overlapped with one another, with the degree of overlap being proportional to the difference between the logarithms of the numerosities. In this way, the performance of the Dehaene-Changeux neural model suggested explanations of the experimental findings from newborns and rats, and it provided a powerful theoretical model for how a mental number line might be realized.

Independent empirical support for a neural mechanism like that hypothesized by Dehaene and Changeux (1993) came from a series of later findings by Nieder, Miller, and colleagues (for a review, see Nieder, 2005). These investigators obtained single-cell recordings of neural activity as an awake animal (a macaque monkey) tracked the number of objects in a set. In a typical task, a set of N objects (dots in various positions, configurations, and sizes)

was presented to the monkey in a sample display, a brief delay was imposed, and then a test display of objects was presented that had either the same number of objects as, or a different number than, the original display. The monkey's task was to respond if the number in the test display matched the number in the sample display.

Like the hypothetical "numerons" in the Dehaene-Changeux model, many neurons were found to be tuned to a particular numerosity (Nieder & Merten, 2007; Nieder & Miller, 2003, 2004). These neurons maintained their numerical selectivity in the face of variation in the position, size, and configuration of the objects in the display (Nieder, Freedman, & Miller, 2002). For example, some neurons were found to show peak firing for a set of one, other neurons peak responding at a set size of two, others at three, and so on, all the way to numbers in the 30s (which were the highest numbers tested). These neurons were involved in numerical memory as well as perception: When the monkey saw a set of four objects that then was hidden and the monkey had to maintain a representation of the number in memory, the fourneurons maintained their activity more than competing neurons did.

Furthermore, the tuning curves of these neurons showed Gaussian variability on a logarithmic scale (as in Fig. 30.1). This means that neurons that peaked when four objects were presented to the monkey would also respond (somewhat less) if three or five objects were presented, and they would respond much less when 1 or 10 objects were presented. Consequently, when the monkey needed to judge whether he was being shown a new number or an old number of objects, accuracy was linked to the difference in logarithms between the two numbers. Thus, collectively, the number-tuned neurons identified by Nieder and colleagues formed a physical basis for the Weber-Fechner law.

These number-tuned neurons were most abundant in the lateral prefrontal cortex (PFC) and in the fundus of the intraparietal sulcus (IPS) (Nieder & Miller, 2004). Neurons in the IPS are thought to code number first, because these neurons required shorter latencies on each trial to become numerosity selective than did PFC neurons (Nieder & Miller, 2004). Based on previous work showing that posterior parietal cortex (PPC) and PFC are functionally interconnected (Cavada & Goldman-Rakic, 1989; Chafee & Goldman-Rakic, 2000; Quintana, Fuster, & Yajeya, 1989), it also seems likely that numerical information first encoded in PPC might

be conveyed directly or indirectly to the PFC, where the signals are amplified and maintained to gain control over behavior. Finally, in a remarkable similarity to the mental number line metaphor, the number-tuned neurons in PPC were found to be so intermingled with neurons that code for line length that number- and length-sensitive neurons were sometimes under the same electrode tip (Tudusciuc & Nieder, 2009)! Thus, the distribution and timing of activation of number-selective neurons suggests that the most likely site for a neural "mental number line" is a frontoparietal circuit.

Could the mental mechanism serving infants' representation of number be the same ones serving monkeys' numerical representations? Because obtaining single-cell recordings requires neurosurgical implantation of electrodes, obtaining directly equivalent measures in human infants is unethical. Therefore, to obtain unobtrusive measures of human infants' neural coding of number, Izard and colleagues (2008) obtained event-related potentials from 3-month-old infants while they were presented with a succession of sets that either changed in number but not object type (e.g., from 4 ducks to 12 ducks) or in object type but not number (e.g., from 4 ducks to 4 balls). The study was designed so that most sets had the same number and type of objects (e.g., repeatedly presenting nonidentical images of 4 ducks). Occasionally, however, a test image appeared that broke this regularity in either number, object identity, or both. The brain response to this disruption was recorded in order to measure the event-related potential (ERP). Across different groups of infants, the numeric changes were 2 versus 3, 4 versus 8, or 4 versus 12.

In all cases, the ERPs revealed that infants' brains detected both types of changes (number and object). To examine the underlying circuitry, the investigators used a source reconstruction method that was based on a model of the infants' cortical folds. This analysis suggested that whereas the infants' left occipitotemporal cortex responded to object novelty, the infants' right parietal cortex responded to numerical novelty, like the corresponding area of the monkeys studied by Nieder.

This method for localizing infants' neural activity is admittedly coarse, due to the relatively poor spatial resolution of electroencephalography (EEG). However, the results were quite similar to those observed in adults and 4-year-olds in studies using more spatially precise functional magnetic resonance imaging (fMRI) methods (Cantlon, Brannon, Carter,

& Pelpfrey, 2006; Piazza, Izard, Pinel, Bihan, & Dehaene, 2004). In these studies, too, regions of the occipitotemporal area reacted to changes in object identity but not to changes in number, whereas posterior parietal regions reacted to changes in number but not to changes in object identity. Both EEG and fMRI data are consistent with the conclusion that number is among the dimensions that are quickly extracted when processing visual stimuli.

Arithmetic in Infancy: Travels on the Mental Number Line

An important property of a number line is that it makes basic addition and subtraction trivial. That is, traveling four spaces forward from four registers the sum of four and four; traveling four spaces back from eight registers the difference between eight and four, and so on.

Therefore, if infants possess a mental number line and encode numerical values of stimuli, they should register sums and differences of numeric quantities (at least approximately). This hypothesis led Wynn (1992) to conduct a series of experiments on infants' arithmetic capacities. To test ability to compute $1 + 1$, for example, Wynn recorded infants' looking times as they watched one object appear to be placed behind an opaque screen and then another object added to it behind the screen. When the screen dropped, seeming to reveal the arithmetically impossible event $1 + 1 = 1$, infants looked longer than when the screen dropped and revealed an outcome consistent with the arithmetically realistic event $1 + 1 = 2$. To be sure that this difference in looking wasn't caused by infants simply preferring to look at one object rather than two, Wynn also examined two objects initially seeming to be placed behind a screen and one object then seeming to be removed. In this situation, infants looked longer when they were shown the arithmetically impossible event of two objects being present after one seemed to be removed, rather than the arithmetically realistic event of one being present. This pattern of findings has since been replicated by other investigators (Koechlin, Dehaene, & Mehler, 1997; Simon, Hespos, & Rochat, 1995).

Testing the idea that infants correctly registered the outcome of the arithmetic operation and were surprised by the arithmetically impossible event, Berger, Tzur, and Posner (2006) collected EEG recordings of infants' brain activity during presentation of correct and incorrect arithmetical operations. In addition to replicating Wynn's finding of

infants looking longer at the seemingly impossible number of objects, infants' brain activity indicated an error detection process in the arithmetically impossible condition. Specifically, the topography and frequency (θ -band effects) of the infants' brain activity were quite similar to what had been found previously in adults when they observe or make responses that they know are wrong, though the error detection signal emerged more slowly in infants than in adults.

An alternative to Wynn's interpretation was that the source of infants' surprise was not violation of their *numeric* expectations but rather a violation of their expectations about objects (Simon, 1997; Uller, Carey, Huntley-Fenner, & Klatt, 1999). In this account, babies used the appearance of the objects to track their individual identities and were surprised when a new individual—not a new number—appeared or went missing. This is a reasonable distinction: When determining whether everyone is present in a small research group, for example, the professor might notice that Tom is missing from the normal attendees of Tom, Dick, and Harry, without bothering to encode how many people were present. As we have seen, infants' abilities to compare sets of four or fewer objects is much greater than would be expected by a mental number line; thus, a subitizing mechanism for tracking small numbers of objects (Kahneman, Treisman, & Gibbs, 1992; Scholl & Leslie, 1999; Trick & Pylyshyn, 1994) might explain infants' error detection on this task.

One way to test whether babies' *numeric* expectations are violated by arithmetically impossible events is to test whether this surprise is also evident when babies witness arithmetically impossible transformations of large sets, where all the individuals couldn't be represented through subitizing (McCrink & Wynn, 2004). Consistent with Wynn's original interpretation that infants were surprised by the numerical outcome, when 9-month-old babies were confronted with arithmetic transformations of large sets (e.g., $5 + 5 = 10$ versus $5 + 5 = 5$, or $10 - 5 = 5$ versus $10 - 5 = 10$), they also looked longer at the arithmetically impossible events than the arithmetically possible ones.

One reason why infants might look longer at the "impossible" outcome is that a logarithmically compressed mental number-line representation supports babies' ability to register approximate sums and differences. If adding n_2 to n_1 involves traveling n_2 spaces forward from n_1 on the mental number line, then babies would arrive at the position

$\log(n_1) + \log(n_2)$, which would be a considerable overestimation of the actual result. Conversely, subtraction through traversing a mental number line would yield a considerable underestimation of the actual results. From this perspective, babies would have no doubt found $5 + 5 = 5$ surprising because they would have experienced it as $\log(5) + \log(5)$ and thus expected to see $\log(25)$!

Consistent with this compressed mental number-line interpretation, McCrink and Wynn (2009) reported that 9-month-old babies' expectations of arithmetic transformations overestimate the results of additive operations and underestimate the results of subtractive operations. Thus, when babies were initially shown a sequence of events equivalent to $6 + 4$, they looked significantly longer when the raised screen revealed 5 objects than when it revealed 20 objects. Similarly, when babies were shown a sequence equivalent to $14 - 4$, they looked significantly longer when the raised screen revealed 20 objects than when it revealed 5. In the same study, the infants looked equally long at 20 objects as 10 at the end of the first example, and looked as long at 5 objects as at 10 at the end of the second example. McCrink and Wynn (2009) explained these findings in terms of "operational momentum." In particular, they suggested that infants moved too far in a positive direction along the mental number line when they added sets of dots and too far in a negative direction when they subtracted sets of dots because they didn't know where to stop their travel along the mental number line.

Another possibility is that the results reflected the infants using logarithmically compressed representations of numerical quantity. In particular, the arithmetically impossible events (e.g., $6 + 4 = 20$; $14 - 4 = 5$) that babies looked at as much as the correct answers were those answers predicted by the view that the infants were mentally representing the problem on a logarithmically compressed mental number line.

Changes Beyond Infancy in Nonsymbolic Numerical Processing

NUMERICAL DISCRIMINATION

Infants' discrimination between sets of objects that differ in number predicts several features of older children's and adults' performance on similar, nonverbal numerosity discrimination tasks. Recall the conclusion that the probability of infants' discrimination was related to the difference between the logarithms of the number of objects in the

two sets. When prevented from counting sets of objects (e.g., through tight time limits or interpolated tasks), older children's and adults' reaction times show the same pattern. For example, the time required for children and adults to select the larger set is proportional to the logarithm of the distance between the numbers of objects in the two sets, both for sets of less than 10 objects (Buckley & Gillman, 1974; Huntley-Fenner & Cannon, 2000) and for sets of 10 or more objects (Birnbaum, 1980; Piazza et al., 2004; Ratcliff, Love, Thompson, & Opfer, 2012).

Another conclusion from infants' performance was that the difference in logarithms needed to discriminate the number of objects in two sets decreases with age. The relation between set size and solution times shows the same pattern in childhood. The time required to discriminate between two numbers that are close together decreases from early childhood to later childhood to adulthood (Ratcliff et al., in press). The "internal Weber fraction," which indicates the difference in ratios needed to reliably discriminate set sizes, has proved useful for quantifying this type of developmental change (Izard et al., 2008; Piazza et al., 2004). Piazza et al. (2010) reported an exponential decline in the internal Weber fraction from an average of 0.34 for kindergarteners, down to 0.25 for 10-year-olds, to 0.15 for adults. Similarly, Halberda and Feigenson (2008) reported internal Weber fractions of 0.38 for kindergarteners and 0.11 for adults. Interestingly, individual differences in Weber fractions within an age group are correlated with the students' math achievement test scores (Halberda, Mazzocco, & Feigenson, 2008). This pattern supports the hypothesis that the precision of the mental number line is foundational to other quantitative skills.

A third conclusion from studies of infants' numerical abilities is that infants' discrimination of very small numbers of objects is consistently higher than would be expected from considering the Weber fractions in isolation. The same is true for older children's and adults' solution times on similar problems. That is, the time required to judge the greater of two sets of three or fewer objects is uniformly low, much lower than would be expected given their ratios alone (Chi & Klahr, 1975; Mandler & Shebo, 1982; Oyama, Kikuchi, & Ichihara, 1981; Trick & Pylyshyn, 1994).

The similarities in performance among infants, older children, and adults suggest that the process of numerosity discrimination is similar, with all

groups accessing the mental number-line representation. This hypothesis is consistent with findings of substantial overlap between the regions in the intraparietal sulcus that are activated by infants, younger children, older children, and adults when comparing numerosities (Castelli, Glaser, & Butterworth, 2006; Piazza et al., 2004; Piazza, Mechelli, Price, & Butterworth, 2006; Piazza, Pinel, Le Bihan, & Dehaene, 2007).

DEVELOPMENT OF NONSYMBOLIC ARITHMETIC BEYOND INFANCY

To examine the development of nonsymbolic arithmetic beyond infancy, McCrink, Dehaene, and Dehaene-Lambertz (2007) showed adults several hundred videos of two successive sets of dots and asked them to approximate their sum or difference by choosing one of two sets of dots. As with infants, adults almost always overshot the correct outcomes on addition problems, as if they had moved “too far” along the mental number line, whereas they almost always undershot the correct outcomes on the subtraction problems, again moving too far on the mental number line in the other direction. In addition, the researchers found that from adults’ modal response, the distribution of other responses tapered off as a function of the ratio of the true and alternative quantities, just as would be predicted by Weber-Fechner’s law. To illustrate the magnitude of this error, the presented subtraction problem of $32 - 16 = 8$ was judged to be correct approximately 60% of the time, which is quite a radical departure from moving along a linearly scaled mental number line.

Development of Symbolic Quantitative Thinking in Childhood and Beyond

In the previous section, we reviewed evidence that a logarithmically-compressed mental number line is a powerful construct for understanding nonsymbolic representations of numerical magnitude, that the construct provides a good analogy to the way that the brain codes numerosity, and that the hypothesis accounts for prominent features of infants’, older children’s, and adults’ nonsymbolic arithmetic capacities.

In this section, we examine the extent to which early-developing, nonsymbolic representations of quantity serve as a foundation for later-developing, symbolic ones. As we will show, on some tasks, the logarithmic function translating symbolic numeric quantities into a subjective representation is preserved from early childhood through adulthood.

We will also see that new ways of thinking about number emerge with age and experience with symbolic expressions of numbers, thereby preparing children to acquire more complex quantitative skills. The new and old numerical representations coexist after the new forms are acquired, with their use varying with the situation and with the child’s numerical experience and mathematical aptitude. Evidence for these ideas comes largely from tasks such as number-line estimation that require participants to translate between alternative quantitative representations, at least one of which is inexact and at least one of which is a symbolically expressed number. Finally, we examine the influence of internal representations of number on uses of those representations in symbolic arithmetic, their relation to overall mathematical achievement, and ways in which research on the mental number line has led to effective instructional interventions.

Numerical Symbols on the Mental Number Line

How do children learn to use the symbols of the decimal system to think about quantities? And how does mastery of this symbol system expand their mathematical capabilities? Early in the learning process, numeric symbols—like the Arabic numerals in Diester and Nieder’s study—are meaningless stimuli for young preschoolers. For example, 2- and 3-year-olds who count flawlessly from 1–10 have no idea that $6 > 4$ and $8 > 6$, regardless of whether the number symbols are spoken or written, nor do children of these ages know how many pennies to give an adult who asks for 4 or more (and very early on, not even how many to give an adult who asks for 1–3) (Le Corre & Carey, 2007; Le Corre, Van de Walle, Brannon, & Carey, 2006; Sarnecka & Carey, 2008).

As young children gain experience with the symbols in a given numerical range and associate them with nonverbal quantities in that range, they initially map them to a logarithmically-compressed mental number line (Berteletti, Lucangeli, Piazza, Dehaene, & Zorzi, 2010; Booth & Siegler, 2006; Geary, Hoard, Nugent, & Byrd-Craven, 2008; Opfer, Thompson, & Furlong, 2010; Siegler & Booth, 2004; Siegler & Opfer, 2003; Thompson & Opfer, 2010). Over a period that typically lasts 1–3 years for a given numerical range (0–10, 0–100, or 0–1000), their mapping changes from a logarithmically compressed form to a linear form, in which subjective and objective numerical values increase

in a 1:1 fashion. Use of linear magnitude representations occurs earliest for the numerals that are most frequent in the environment, that is, the smallest whole numbers (Berteletti, Lucangeli, Piazza, Dehaene, & Zorzi, 2010), and it gradually is applied to increasingly large numbers (Siegler, Thompson, & Opfer, 2009). In this section, we first examine this process of learning representations of symbolic numbers as it applies to numerical magnitude comparison and then as it applies to number-line estimation; after this, we examine the consequences of individual differences in these numerical representation processes and interventions that promote the growth of linear representations.

SYMBOLIC NUMBER COMPARISON

When 3-year-olds from middle-income backgrounds or 4- and 5-year-olds from low-income backgrounds are presented numerical magnitude comparison problems involving the numbers 1–9 in symbolic form, their performance is near chance (Ramani & Siegler, 2008; Siegler & Robinson, 1982). In contrast, by kindergarten or first grade, most children are highly accurate on these numerical comparison problems with Arabic numerals. Their solution times, and those of older children and adults, are subject to the same Weber-Fechner law that characterizes numerical comparisons of nonsymbolic stimuli, such as dot arrays (Buckley & Gillman, 1974; Moyer & Landauer, 1967; Sekuler & Mierkiewicz, 1977), with the time required to select the larger of two numbers being inversely proportional to the logarithm of the distance between the numbers being compared. Thus, the time required to select the larger of 3 and 5 is less than the time required to compare 5 and 7, and the time required to select the larger of 30 and 50 is less than the time required to compare 50 and 70 (Dehaene, 1989). This finding suggests that Arabic numerals automatically activate the same mental number-line representation that encodes nonsymbolic numerosity (as in Fig. 30.1).

Consistent with this idea, fMRI studies have revealed that even very brief presentations of number symbols evoke number-related activation in the intraparietal cortex of educated adults (Naccache & Dehaene, 2001). The systems for representing nonsymbolic and symbolic numbers appear to be closely linked. In addition to the physical overlap, if not identity, between the systems, habituation of one leads to habituation of the other. Thus, after adaptation to 17, 18, or 19 dots, dishabituation (as

measured by fMRI activations) is observed when the Arabic numeral 50 is presented, but not when the numeral 20 is presented (Piazza et al., 2007). These results suggest that—at least in educated adults—populations of neurons in parietal and prefrontal cortex are activated by both nonsymbolic numerosities and by number symbols. Put another way, the mental number line appears to link symbolic and nonsymbolic numerical representations.

Given that the invention of Arabic numerals only dates back a few thousand years (Ifrah, 2000), it is implausible that the brain evolved specifically to handle them. More likely, the link between the two systems arises through learning during childhood. To investigate how this learning process might occur, Diester and Nieder (2007) trained two monkeys to associate Arabic numerals with the numerosity of multidot displays. After the training, a large proportion of PFC neurons encoded numerical values, irrespective of whether the numbers had been presented to the monkeys in the form of dots or Arabic numerals. Moreover, these neurons exhibited similar tuning functions, with activity falling with increasing numerical distance from the neuron's "preferred" numerical value. Over the course of training, numeral-numerosity associations progressively shifted from the PFC to the IPS.

A similar pattern may exist in human development (Ansari et al., 2005; Houdé, Rossi, Lubin, & Joliot, 2010; Rivera, Reiss, Eckert, & Menon, 2005). The distance effect observed in monkey's PFC neurons has also been observed in the PFC of preschool children. Moreover, with age and experience, neural activation in response to numerical stimuli shifts to posterior parietal areas, particularly in the left hemisphere, much as the numeral-numerosity associations shifted to IPS during the later stages of training among the monkeys studied by Diester and Nieder (2007). Together, these results suggest that the PFC might be the first cortical area to link numerals to the mental number line, and that with age and experience, the IPS is increasingly involved in linking numerals with the numerosities for which they stand.

NUMBER-LINE ESTIMATION

An alternative method that has been developed to examine the development of representations of numerical magnitudes is the number-line estimation task (Siegler & Opfer, 2003). On this task, participants are shown a blank line flanked by a number at each end (e.g., 0 and 1,000) and asked

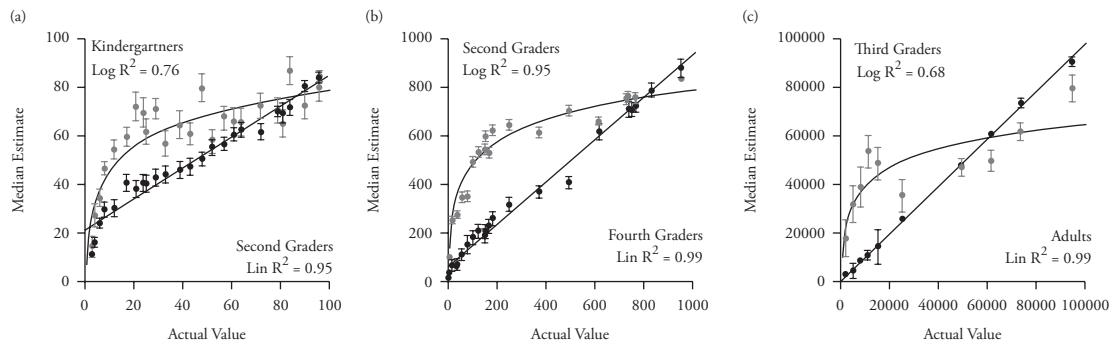


Fig. 30.2 Development of number-line estimation across three scales.

where a third number (e.g., 150) would fall on the line. This task is particularly revealing about representations of numerical magnitude because it transparently reflects the ratio characteristics of the number system. Just as 150 is twice as large as 75, the distance of the estimated position of 150 from 0 should be twice as great as the distance of the estimated position of 75 from 0. More generally, estimated magnitude (y) should increase linearly with actual magnitude (x), with a slope of 1.00, as in the equation $y = x$.

Across a number of cross-sectional studies using this number-line estimation task (Booth & Siegler, 2006; Laski & Siegler, 2007; Opfer & DeVries, 2008; Opfer & Siegler, 2007; Siegler & Booth, 2004; Siegler & Opfer, 2003; Thompson & Opfer, 2008, 2010), a systematic difference between younger and older children's estimates has been evident: Younger children's estimates of numerical magnitude typically follow Fechner's law ($y = k \times \ln x$) and increase logarithmically with actual value (Fig. 30.2). In contrast, older children's estimates for the same range of numbers increase linearly with actual value.

This developmental sequence emerges at different ages with different numerical ranges (Fig. 30.2). It occurs between preschool and kindergarten for the 0–10 range, between kindergarten and second grade for the 0–100 range, between second and fourth grade for the 0–1,000 range, and between third and sixth grade for the 0–100,000 range (Berteletti, Lucangeli, Piazza, Dehaene, & Zorzi, 2010; Opfer & Siegler, 2007; Siegler & Booth, 2004; Thompson & Opfer, 2010). Thus, as shown in Figure 30.2A, on the 0–100 number-line estimation task, the logarithmic function fit kindergartners' estimates better than did the linear function, but the linear function fit second graders' estimates better than did the logarithmic function. The same transition occurs roughly

a year later for children with mathematical learning difficulties (Geary, Hoard, Byrd-Craven, Nugent, & Numtee, 2007). The timing of the changes corresponds to the periods when children are gaining extensive exposure to the numerical ranges: through counting during preschool for numbers up to 10, through addition and subtraction between kindergarten and second grade for numbers through 100, and through all four arithmetic operations in the remainder of elementary school.

How might these changes occur? One clue came from the estimates of sixth graders and adults in Siegler and Opfer (2003). The median estimates of both groups increased linearly with numeric value, but the variability of estimates for both was smallest near the *quartiles* on the number line, as if these older children and adults mapped numbers like 0, 250, 500, 750, and 1,000 to 0, 1, 2, 3, and 4—that is, numbers that other studies indicated they already represent linearly (Chi & Klahr, 1975; Siegler & McGilly, 1989; Siegler & Robinson, 1982). This observation suggested that children might map early-developing linear representations of numerical magnitudes to guide their learning about the magnitudes of less familiar, larger numerals, a process that might be aided through children learning about fractions and percentages in third through fifth grade.

To test this idea about how the logarithmic-to-linear shift occurs, we conducted three microgenetic studies that allowed us to examine changes in numerical estimation on a trial-to-trial basis (Opfer & Siegler, 2007; Opfer & Thompson, 2008; Thompson & Opfer, 2008). In these studies, we used an experimental manipulation to help students who initially used logarithmic representations to adopt a linear representation. Specifically, we provided children with feedback on their estimates of numbers

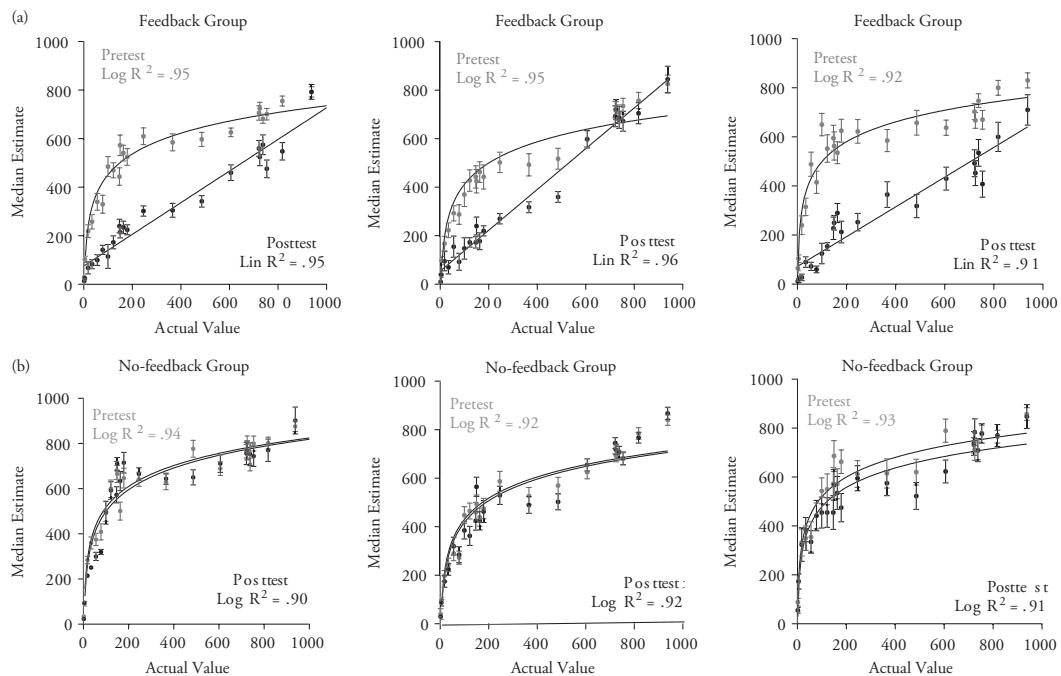


Fig. 30.3 Estimation of numeric magnitudes in response to limited feedback. (a) Across three studies (L-R: Opfer & Siegler, 2007; Opfer & Thompson, 2008; Thompson & Opfer, 2008), logarithmic-to-linear shifts in estimation occurred from pretest (gray circles) to posttest (black circles) when children were given feedback on their estimates for a few numbers around 150. (b) Across the same studies, very little change in logarithmic estimation patterns occurred when children were not given feedback on their estimates.

around 150, the point where the logarithmic and linear functions that pass through 0 and 1,000 are most discrepant. The idea was that this feedback would be highly salient, due to it indicating that the children's estimates were far from correct, and that the experience would lead children to draw an analogy between what the feedback indicated about the correct placement of 150 on a 0–1,000 number line and their existing knowledge of the correct placement of 15 on a 0–100 number line. After receiving feedback on their estimates of numbers around 150, second graders in all three studies provided estimates that increased linearly with actual value (Fig. 30.3A). This representational change was evident in differences in the median estimates of the treatment groups on pretest and posttest, with pretest estimates being best fit by logarithmic functions and posttest estimates being best fit by linear functions. Providing feedback regarding correct placements at other numerical values, such as 725, also increased the linearity of estimates, but change occurred more slowly and to a somewhat lesser extent (Opfer & Siegler, 2007). Consistent with the interpretation that this change came about through

a process of analogy, when children were given an opportunity to compare (1) the placement of 15 cherries on a line flanked by 0 and 100 cherries, and (2) the placement of 1,500 on a line flanked by 0 and 10,000, children quickly apprehend the similarity of the two situations and use it to improve their estimates of the larger numbers (Thompson & Opfer, 2010).

As is evident in Figure 30.3, second graders' newly adopted linear representation spanned the entire 0–1,000 range and was not simply a "local fix" to the numbers near 150. This effect across the entire numerical range was consistent with the view that the numerical magnitude representation has psychological reality as a coherent unit, rather than simply being a convenient way of summarizing data. Moreover, like an analogical insight (e.g., Gick & Holyoak, 1983; see Holyoak, Chapter 13), the shift occurred abruptly, much more abruptly than what had been observed in studies of transitions in related areas such as addition and numerical insight problems (e.g., Siegler & Jenkins, 1989; Siegler & Stern, 1998). Often, feedback on a single trial was sufficient to yield a shift from a logarithmic to a linear

pattern of estimates (Opfer & Siegler, 2007; Opfer & Thompson, 2008; Thompson & Opfer, 2008).

The finding of new symbolic representations arising through analogies to representations of similar relations in simpler contexts is widely evident in cognitive development (for a review, see Gentner, 2010). Such analogy-based representational change is important, because it suggests a potential solution to a problem that is endemic to the mental number-line representation discovered by Nieder and colleagues. The problem is that it seems as though one would simply run out of “numerons” to represent large numbers. Do people and other animals possess neurons that fire more in response to seeing 1,582 dots than in response to 1,583? However, if learners can map small numbers to large ones, they would gain the ability to use the relations among small number representations to represent relations among larger magnitudes (as when adults mapped 250-500-750-1,000 to 1-2-3-4 or $\frac{1}{4}$, $\frac{1}{2}$, $\frac{3}{4}$, 1 or to 25%, 50%, 75%, and 100%). If this occurs at the neural level (and in some sense it must), then having a limited number of numerons imposes no inherent limit on the number of numeric magnitudes that could be represented.

MEASUREMENT AND NUMEROSITY ESTIMATION

To examine the generality of the developmental transition revealed by the number-line task, several studies have tested whether similar changes are evident in estimates of line lengths and numbers of discrete objects (measurement estimation and numerosity estimation) (Booth & Siegler, 2006; Laski & Siegler, 2007; Thompson & Siegler, 2010). On the measurement estimation task, children saw an extremely short line, labeled “1 zip;” a long line, labeled “1,000 zips;” and a number indicating the length of a line (in zips) that should be drawn. Children drew a line to approximate the desired length, then were asked to draw a line of a different number of zips, and so on. On the numerosity estimation task, children saw a computer screen that depicted one box with 0 dots, one with 1,000 dots, and a third, initially empty, box that could be filled with the desired number of dots by placing the cursor in the “increase” box or the “decrease” box and holding down the mouse until the desired number of dots was reached (the time limit was too short for the children to count the dots).

On each task, the same logarithmic-to-linear shift that had been observed with number-line estimation was observed (Booth & Siegler, 2006; Thompson

& Siegler, 2010). On the measurement estimation task in Booth and Siegler (2006), for example, the variance accounted for by the best-fitting linear function increased with age (from 85% to 98%), whereas the variance accounted for by the best-fitting logarithmic function decreased with age (from 91% to 74%). Furthermore, individual second graders’ measurement estimates were more likely to be better fit by the logarithmic function than by the linear function (70% vs. 30%), whereas fourth graders’ estimates were more likely to be better fit by the linear function than by the logarithmic function (78% vs. 22%). Similarly, the percentage of children for whom the linear function provided the better fit on the numerosity estimation task increased from 43% to 82%, whereas the percentage of children for whom the logarithmic function provided the better fit decreased from 57% to 18%.

NUMBER CATEGORIZATION

To examine whether the logarithmic-to-linear transition extended beyond estimation tasks, Laski and Siegler (2007) presented 5- to 8-year-olds with a numerical categorization task. Children were told that 1 was a “really small number” and that 100 was a “really big number” and then were asked to categorize numbers between 1 and 100 as “really small,” “small,” “medium,” “big,” or “really big.” Each child’s categorization of each number was assigned a numerical value ranging from 1 for the “really small” category to 5 for the “really big” category. Then, the mean value for the categorizations of each number was computed, and the fit of linear and logarithmic functions to the mean categorization scores for the full set of numbers was calculated.

Kindergartners’ mean categorizations of the numbers were better predicted by the best-fitting logarithmic function than by the best-fitting linear function. In contrast, second graders’ mean categorizations were (nonsignificantly) better predicted by the best-fitting linear function than by the best-fitting logarithmic function. The same change was evident for the number-line task. Moreover, the linearity of individual children’s number-line estimation and categorization patterns was highly correlated, providing additional evidence for the generality of the logarithmic-to-linear transition in representations of numerical magnitudes.

If number-line estimation and categorization reflect the same underlying numerical representation, then experience that leads to improved number-line estimation might transfer to numerical

categorization. To test this hypothesis, Opfer and Thompson (2008) presented kindergartners who initially produced number-line estimates that were more logarithmic than linear with feedback designed to improve the linearity of number-line estimates and tested both number-line estimation and categorization. As expected, after the feedback experience, the linear function provided a better fit to the mean category judgments of children who received feedback on their number-line judgments than for those who did not receive such feedback. Thus, the change from a logarithmic to a linear representation on the number-line task extended to the categorization task, even without any feedback on that task.

Symbolic Arithmetic in Childhood

Linear representations of numerical magnitudes also are important for symbolic arithmetic abilities, both for approximating answers and for learning exact sums. One source of evidence regarding approximations to correct sums comes from Gilmore, McCarthy, and Spelke's (2007) study of preschoolers' estimates of answers to arithmetic problems that they had not yet encountered in school. The investigators presented 5- and 6-year-olds with problems such as "Sarah has 21 candies, she gets 30 more, John has 34 candies—who has more?" To insure that the preschoolers understood the symbols that were being used, the problems were simultaneously presented both orally—as spoken numerals—and in writing, as Arabic numerals.

Despite the fact that the preschoolers had received no training with numbers of that size, they spontaneously performed better than chance (60%–75%). This was true regardless of their socioeconomic origins. Performance was still approximate, however, and depended on the ratio of the two sums that the children were choosing between, a signature of the Weber-Fechner law.

Other evidence for the relation between the linearity of numerical magnitude representations and arithmetic knowledge comes from positive relations between the linearity of first through fourth graders' estimates on number line, measurement, and numerosity tasks on the one hand and the accuracy of their estimates of answers to two digit plus two digit addition problems on the other (Booth & Siegler, 2006). Yet other correlational evidence comes from positive relations in first graders' linearity of number-line estimates and the number of single digit addition problems that the children

answered correctly (Booth & Siegler, 2008; Geary et al., 2007).

In addition to this correlational evidence, linear representations of numerical magnitude also play a causal role in arithmetic learning. Booth and Siegler (2008) pretested firstgraders' number-line estimation and retrieval of answers to 13 addition problems, ranging in difficulty from $1 + 4$ to $49 + 43$. Then, children were trained on the easiest 2-digit + 2-digit problems that they had answered incorrectly on the pretest. All children were presented with each of these addition problems three times, with feedback regarding the correct answer being provided after each presentation.

A randomly chosen half of the children were also presented with analog linear representations of the addends and sum. This manipulation was intended to inculcate a linear representation of the numbers in the addition problems and ideally of numbers in the 0–100 range more generally. Children in this experimental condition saw a number line with 0 at one end and 100 at the other, then saw the first addend represented by a red bar just above the line, then the second addend represented by a blue bar just below the line, and then the sum represented by a purple bar straddling the line. Thus, if the problem was $43 + 49$, the red bar would be 43% of the number line's length, the blue bar 49% of its length, and the purple bar 92% of its length. The logic was that seeing the linear representations of the addends and sums along the number line would allow children to encode the numerical magnitudes more accurately and thus help them retrieve the answers to the problems.

Presentation of the analog representations of the addends and sum along the number line was causally related to arithmetic learning; it increased the number of addition problems correctly recalled on the posttest and also improved the linearity of the children's number-line estimates. Moreover, the effect of the experimental manipulation was even stronger for measures of the closeness of addition errors to the correct sum than for the number of correct sums, supporting the view that activating the linear representation was the means through which the experimental manipulation produced its effect. If this mechanism were not involved, why else would children who were presented with the analog representations of the addends and sums increasingly advance incorrect answers that were close to the sum and decreasingly produce answers that were far from it? Similarly, randomly assigning

preschoolers to play a linear numerical board game, rather than a color board game, has been found to increase preschoolers' learning of answers to arithmetic problems (Siegler & Ramani, 2009). Thus, linear magnitude representations are both causally and relationally related to arithmetic learning.

RELATIONS BETWEEN NUMERICAL MAGNITUDE REPRESENTATIONS AND OVERALL MATHEMATICS PROFICIENCY

Both nonsymbolic and symbolic numerical magnitude representations have been found to be related to standardized mathematics achievement test performance. Halberda et al. (2008) found that accuracy of nonsymbolic number comparison was positively related to achievement test scores in mathematics, but not to other domains of the school curriculum. Similarly, Holloway and Ansari (2008) found that individual differences in the distance effect during symbolic number comparison among children aged 6 to 8 years is related to mathematics achievement but not reading achievement.

Moreover, the linearity of number line, measurement, and numerosity estimation, as well as of numerical categorization, have been related to mathematics achievement test performance (Booth & Siegler, 2006, 2008; Laski & Siegler, 2007); and children with mathematical learning difficulties, as defined by low mathematics achievement test scores, often generate logarithmic patterns of estimates for several years beyond the time when other students have adopted linear representations for the same numerical range (Geary et al., 2007; Geary et al., 2008). Thus, linear representations of numerical magnitudes seem related to general as well as specific numerical competencies.

THEORETICALLY-BASED EDUCATIONAL INTERVENTIONS

Research on the centrality of the mental number line in numerical knowledge suggested an educational intervention that has proved highly effective with low-income preschoolers. The intervention began with the question, "How do children develop a linear representation of numerical magnitudes?" Experience with counting likely contributes, but such experience is insufficient for children to construct linear representations of numerical magnitudes, as indicated by the previously described dissociation between counting knowledge and magnitude comparison of the numbers that can be counted.

One common activity that was hypothesized by Siegler and Booth (2004) to help children generate linear representations is playing linear, number board games—that is, board games with linearly arranged, consecutively numbered, equal-size spaces (e.g., *Chutes and Ladders*.) These board games provide multiple cues to numbers' magnitudes. The greater the number in a square, the greater: (a) the distance that the child has moved the token, (b) the number of discrete moves of the token the child has made, (c) the number of number names the child has spoken, (d) the number of number names the child has heard, and (e) the amount of time since the game began. Thus, children playing the game have the opportunity to relate the number in each square to the time, distance, and number of manual and vocal actions required to reach that number.

To test whether playing number board games promotes number sense, Ramani and Siegler (2008; Siegler & Ramani, 2008, 2009) randomly assigned 4- and 5-year-olds from low-income backgrounds to play either a number board game or a color board game. At the beginning of each session, children in the number board condition were told that on each turn, they would spin a spinner that would point to "1" or "2," that they should move their token that number of spaces, and that the first player to reach the end would win. Children in the color board condition were told that on each turn, they would spin a spinner that could point to different colors, that they should move their token to the nearest square with the same color as the one to which the spinner pointed, and that the first player to reach the end would win. The experimenter also told children to say the numbers (colors) on the spaces through which they moved. Thus, children in the number board group who were on a 3 and spun a 2 would say, "4, 5" as they moved their token. Children in the color board group who were on green and spun a "blue" would say "purple, blue." If a child erred or could not name the numbers or colors, the experimenter correctly named them and then had the child repeat the names while moving the token. The preschoolers played the number game or the color game about 20 times over four 15-minute sessions within a 2-week period; each game lasted about 3 minutes.

Playing the number board game led to dramatic improvements in the low-income preschoolers' number-line estimates. Accuracy of number-line estimation, magnitude comparison, counting, and numeral identification increased from pretest to posttest among children who played the number

board game. Gains remained present on a follow-up 9 weeks later. In contrast, there was no change in the accuracy of estimates of children who played the color board game (Ramani & Siegler, 2008; Siegler & Ramani, 2009). Playing the game also improved the children's learning of subsequently presented addition problems (Siegler & Ramani, 2009).

Conclusions and Future Directions

Quantitative thinking, ranging from the ability to choose the greater of two sets of a few objects to the ability to project arithmetic transformations with very large numbers, is central to the lives of humans and other animals. In this chapter, we have argued that despite quantitative thinking playing important roles in the lives of both humans and other animals, there is also a fundamental distinction present between two kinds of quantitative thinking—nonsymbolic and symbolic—and that this distinction is essential to understanding what develops in human quantitative thinking.

On nonsymbolic quantitative tasks, similarities between human and nonhuman quantitative thinking pervade many levels of analysis. In a wide range of species and age groups (*1*) there is a capacity to mentally represent and compare the approximate number of objects or events in sets; (*2*) mental representations of nonsymbolic numeric quantities are noisy and increase logarithmically with actual quantity, leading to the speed and accuracy of comparison depending on the ratio of the two numbers, and (*3*) the neural correlates of nonsymbolic number representation are distributed in regions of frontoparietal cortex that overlap greatly across species and age groups from infants to adults. A common cognitive mechanism—a logarithmically compressed “mental number line”—also appears to underlie thinking about nonsymbolic quantities in people and other animals.

Nonsymbolic quantitative thinking, however, has important limitations. Nonsymbolic representations of quantitative properties of sets greater than three or four objects are inexact, making impossible activities such as economic transactions that require precision (Furlong & Opfer, 2009). Lack of symbols also makes it impossible to communicate numerical properties to other minds (e.g., that there are 12 sheep in one herd but 18 in another). Furthermore, without symbols, it is impossible to preserve numerical information over time, as when recording tallies of livestock in a ledger, and it is impossible to track small but important changes, for example, gradual increments in the size of a herd. Moreover, mentally

adding or subtracting nonsymbolic quantities of any substantial size yields extremely noisy estimates that preclude many characteristic human activities, such as deciding on a fair exchange involving numerous specific items. Thus, while the capacity to represent and compare nonsymbolic quantities is important, and provides a basis for symbolic numerical capabilities, the capacity to represent symbolic quantities offers crucial cognitive advantages.

This theoretical contrast between nonsymbolic and symbolic quantitative abilities provides a useful way to approach development of quantitative thinking. Numeric symbols were invented relatively recently in human history, making it implausible that the brain evolved specifically to handle them; they must be learned through prolonged interaction with the environment. Consistent with this perspective, nonsymbolic quantitative thinking emerges early in infancy and shows little variation among children of different cultural backgrounds. In contrast, symbolic quantitative thinking emerges much later, typically in preschool among middle and upper class groups in developed societies and even later among less affluent groups in those societies. Indeed, numerical symbols are used only for a few small whole numbers among children in adults in less developed societies, such as some indigenous Amazonian populations (Dehaene, Izard, Spelke, & Pica, 2008). For all of these reasons and many more, the distinction between nonsymbolic and symbolic quantitative thinking is fundamental to describing and understanding the development of quantitative thinking.

As we have seen, children's learning of the decimal system and the meaning of the symbols within it occurs over many years through substantial observation of adults and older children, direct instruction from teachers and parents, extensive practice, and surmounting misunderstandings and forming new understandings. Early in the learning process, numerical symbols are meaningless stimuli for young preschoolers. For example, 2- and 3-year-olds who count flawlessly from 1–10 have no idea that 6>4, nor do children of these ages know how many pennies to give an adult who asks for 4 or more. As young children gain experience with the symbols in a given numerical range and associate them with nonverbal quantities in that range, they initially map them to a logarithmically compressed mental number line. Over a period that typically lasts 1–3 years for a given numerical range (0–10, 0–100, or 0–1,000), children's mapping of symbolically expressed numbers to nonverbal representations changes from a

logarithmically compressed form to a linear form. In the linear representation, subjective and objective numerical values increase in a 1:1 fashion. Use of linear magnitude representations occurs earliest for the numerals that are most frequent in the environment, that is, the smallest whole numbers, and is gradually extended to increasingly large numbers.

The logarithmic-to-linear shift in children's representations of symbolic quantities expands children's quantitative thinking profoundly. It improves (1) children's ability to estimate the positions of numbers on number lines, (2) to estimate the measurements of continuous and discrete quantities, (3) to categorize numbers according to size, (4) to remember numbers they have encountered, and (5) to estimate and learn the answers to arithmetic problems. These abilities are crucial for learning mathematics, resulting in the use of linear representations of number being highly correlated with skill at arithmetic and overall mathematics achievement. Especially important, educational interventions aimed at inculcating linear representations have broad and sustained effects on mathematics performance and learning. In contrast, interventions aimed at improving nonsymbolic quantitative abilities (e.g., Räsänen, Salminen, Wilson, Aunio, & Dehaene, 2009) have proved less effective in leading to positive educational outcomes.

A final contrast between development of nonsymbolic and symbolic quantitative thinking is worth highlighting as an important direction for future research. Although much is known about the neural substrate of nonsymbolic quantitative abilities, and a reasonable amount is known about its development, little is known about several remaining issues concerning symbolic numerical capabilities: (1) What are the neural substrates for acquiring the ability to think about symbolic numerical quantities? (2) How do the neural correlates of symbolic number representation change with age and experience? (3) How does a limited set of neurons represent the magnitudes of a much larger set of symbolic numeric quantities, including those expressed by integers, fractions, and ordinals? and (4) What mechanisms for representing symbolic number are affected by the various developmental disorders (e.g., William's syndrome) that are marked by difficulties with mathematics?

References

- Ansari, D., Garcia, N., Lucas, E., Hamon, K., & Dhital, B. (2005). Neural correlates of symbolic number processing in children and adults. *Neuroreport*, 16, 1769–1773.
- Antell, S., & Keating, D. (1983). Perception of numerical invariance in neonates. *Child Development*, 54(3), 695–701.
- Berger, A., Tzur, G., & Posner, M. (2006). Infant brains detect arithmetic errors. *Proceedings of the National Academy of Sciences USA*, 103(33), 12649.
- Berteletti, I., Lucangeli, D., Piazza, M., Dehaene, S., & Zorzi, M. (2010). Numerical estimation in preschoolers. *Developmental Psychology*, 46(2), 545–551.
- Bijeljac-Babic, R., Bertoncini, J., & Mehler, J. (1993). How do 4-day-old infants categorize multisyllabic utterances? *Developmental Psychology*, 29, 711–721.
- Birnbaum, M. (1980). Comparison of two theories of "ratio" and "difference" judgments. *Journal of Experimental Psychology: General*, 109(3), 304–319.
- Booth, J., & Siegler, R. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology*, 41(6), 189–201.
- Booth, J., & Siegler, R. (2008). Numerical magnitude representations influence arithmetic learning. *Child Development*, 79(4), 1016–1031.
- Boysen, S., & Capaldi, E. J. (1993). *The development of numerical competence: animal and human models*. Hillsdale, NJ: Erlbaum.
- Brannon, E., Abbott, S., & Lutz, D. (2004). Number bias for the discrimination of large visual sets in infancy. *Cognition*, 93, B59–B68.
- Brannon, E., Libertus, M., Meck, W., & Woldorff, M. (2008). Electrophysiological measures of time processing in infant and adult brains: Weber's law holds. *Journal of Cognitive Neuroscience*, 20, 193–203.
- Brannon, E., Lutz, D., & Cordes, S. (2006). The development of area discrimination and its implications for number representation in infancy. *Developmental Science*, 9, F59–F64.
- Buckley, P. B., & Gillman, C. B. (1974). Comparisons of digits and dot patterns. *Journal of Experimental Psychology*, 103, 1131–1136.
- Butterworth, B., Reeve, R., Reynolds, F., & Lloyd, D. (2008). Numerical thought with and without words: Evidence from indigenous Australian children. *Proceedings of the National Academy of Sciences USA*, 105, 13179–13184.
- Cantlon, J., Brannon, E., Carter, E., & Pelpfrey, K. (2006). Functional imaging of numerical processing in adults and 4-y-old children. *PLoS Biology*, 4 (5), e125.
- Castelli, F., Glaser, D. E., Butterworth, B. (2006). Discrete and analogue quantity processing in the parietal lobe: A functional MRI study. *Proceedings of the National Academy of Sciences USA*, 103, 4693–4698.
- Cavada, C., & Goldman-Rakic, P. S. (1989). Posterior parietal cortex in rhesus monkey: II. Evidence for segregated corticocortical networks linking sensory and limbic areas with the frontal lobe. *Journal of Comparative Neurology*, 287, 422–445.
- Chafee, M. V., & Goldman-Rakic, P. S. (2000). Inactivation of parietal and prefrontal cortex reveals interdependence of neural activity during memory-guided saccades. *Journal of Neurophysiology*, 83, 1550–1566.
- Chi, M. T. H., & Klahr, D. (1975). Span and rate of apprehension in children and adults. *Journal of Experimental Child Psychology*, 19, 434–439.
- Clearfield, M., & Mix, K. (1999). Number versus contour length in infants' discrimination of small visual sets. *Psychological Science*, 10, 408–411.

- Clearfield, M., & Mix, K. (2001). Amount versus number: Infants' use of area and contour length to discriminate small sets. *Journal of Cognition and Development*, 2, 243–260.
- Cordes, S., & Brannon, E. (2008a). Quantitative competencies in infancy. *Developmental Science*, 11(6), 803–808.
- Cordes, S., & Brannon, E. (2008b). The difficulties of representing continuous extent in infancy: Using number is just easier. *Child Development*, 79(2), 476–489.
- Cordes, S., & Brannon, E. (2009). The relative salience of discrete and continuous quantity in young infants. *Developmental Science*, 12(3), 453–463.
- Davis, H. (1993). Numerical competence in animals: A conservative view. In S. T. Boysen & E. J. Capaldi (Eds.), *The development of numerical competence: Animal and human models* (pp. 109–125). Hillsdale, NJ: Erlbaum.
- Dehaene, S. (1989). The psychophysics of numerical comparison: A reexamination of apparently incompatible data. *Perception and Psychophysics*, 45(6), 557–566.
- Dehaene, S. (1992). Varieties of numerical abilities. *Cognition*, 44, 1–42.
- Dehaene, S., & Changeux, J.-P. (1993). Development of elementary numerical abilities: A neuronal model. *Journal of Cognitive Neuroscience*, 5(4), 390–407.
- Dehaene, S., Izard, V., Pica, P., & Spelke, E. S. (2008). Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigenous cultures. *Science*, 320, 1217–1220.
- Diester, I., & Nieder, A. (2007). Semantic associations between signs and numerical categories in the prefrontal cortex. *PLoS Biology*, 5(11), e294.
- Fantz, R., & Fagan, J. (1975). Visual attention to size and number of pattern details by term and preterm infants during the first six months. *Child Development*, 46, 3–18.
- Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, 8(7), 307–314.
- Furlong, E. E., & Opfer, J. E. (2009). Cognitive constraints on how economic rewards affect cooperation. *Psychological Science*, 20, 11–16.
- Gallistel, C. R. (1993). A conceptual framework for the study of numerical estimation and arithmetic reasoning in animals. In S. T. Boysen & E. J. Capaldi (Eds.), *Development of numerical competence: Animal and human models* (pp. 211–224). Hillsdale, NJ: Erlbaum.
- Gao, F., Levine, S., & Huttenlocher, J. (2000). What do infants know about continuous quantity? *Journal of Experimental Child Psychology*, 77, 20–29.
- Geary, D., Hoard, M., Byrd-Craven, J., Nugent, L., & Numtee, C. (2007). Cognitive mechanisms underlying achievement deficits in children with mathematical learning disability. *Child Development*, 78(4), 1343–59.
- Geary, D. C., Hoard, M. K., Nugent, L., & Byrd-Craven, J. (2008). Development of number line representations in children with mathematical learning disability. *Developmental Neuropsychology*, 33, 277–299.
- Gelman, R., & Gallistel, C. R. (1978). *The child's understanding of number*. Cambridge, MA: Harvard University Press.
- Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science: A Multidisciplinary Journal*, 34, 752–775.
- Gick, M., & Holyoak, K. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, 15, 1–38.
- Gilmore, C. K., McCarthy, S. E., & Spelke, E. S. (2007). Symbolic arithmetic knowledge without instruction. *Nature*, 447, 589–591.
- Gordon, P. (2004). Numerical cognition without words: Evidence from Amazonia. *Science*, 306, 496–499.
- Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the "number sense": The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental Psychology*, 44(5), 1457–1465.
- Halberda, J., Mazzocco, M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, 455(7213), 665–668.
- Huntley-Fenner, G., & Cannon, E. (2000). Preschoolers' magnitude comparisons are mediated by a preverbal analog mechanism. *Psychological Science*, 11, 147–152.
- Holloway, I. D., & Ansari, D. (2008). Domain-specific and domain-general changes in children's development of number comparison. *Developmental Science*, 11, 644–649.
- Houdé, O., Rossi, S., Lubin, A., & Joliot, M. (2010). Mapping numerical processing, reading, and executive functions in the developing brain: An fMRI meta-analysis of 52 studies including 842 children. *Developmental Science*, 13, 876–885.
- Ifrah, G. (2000). *The universal history of numbers: From prehistory to the invention of the computer*. New York: Wiley.
- Izard, V., Dehaene-Lambertz, G., & Dehaene, S. (2008). Distinct cerebral pathways for object identity and number in human infants. *PLoS Biology*, 6(2), e11.
- Izard, V., Sann, C., Spelke, E. S., & Streri, A. (2009). Newborn infants perceive abstract numbers. *Proceedings of the National Academy of Sciences USA*, 106, 10382–10385.
- Jordan, K. E., & Brannon, E. (2006a). A common representational system governed by Weber's law: Nonverbal numerical similarity judgments in 6-year-olds and rhesus macaques. *Journal of Experimental Child Psychology*, 95(3), 215–229.
- Jordan, K. E., & Brannon, E. (2006b). The multisensory representation of number in infancy. *Proceedings of the National Academy of Sciences USA*, 103, 3486–3489.
- Jordan, K. E., Suanda, S. H., & Brannon, E. M. (2008). Intersensory redundancy accelerates preverbal numerical competence. *Cognition*, 108, 210–221.
- Kahneman, D., Treisman, A., & Gibbs, B. (1992). The reviewing of object-files: Object specific integration of information. *Cognitive Psychology*, 24, 175–219.
- Kobayashi, T., Hiraki, K., & Hasegawa, T. (2005). Auditory-visual intermodal matching of small numerosities in 6-month-old infants. *Developmental Science*, 8, 409–419.
- Koechlin, E., Dehaene, S., & Mehler, J. (1997). Numerical transformations in five-month-old human infants. *Mathematical Cognition*, 3, 89–104.
- Laski, E., & Siegler, R. (2007). Is 27 a big number? Correlational and causal connections among numerical categorization, number line estimation, and numerical magnitude comparison. *Child Development*, 78, 1723–1743.
- Le Corre, M., & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*, 105, 395–438.
- Le Corre, M., Van de Walle, G., Brannon, E. M., & Carey, S. (2006). Re-visiting the competence/performance debate in the acquisition of the counting principles. *Cognitive Psychology*, 52, 130–169.
- Lipton, J., & Spelke, E. S. (2003). Origins of number sense: Large number discrimination in human infants. *Psychological Science*, 14, 396–401.

- Lipton, J., & Spelke, E. S. (2004). Discrimination of large and small numerosities by human infants. *Infancy*, 5, 271–290.
- Loureiro, S., & Longo, M. (2010). General magnitude representation in human infants. *Psychological Science*, 21, 873–881.
- Mandler, G., & Shebo, B. J. (1982). Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General*, 111, 1–22.
- McComb, K., Packer, C., & Pusey, A. (1994). Roaring and numerical assessment in contests between groups of female lions, *Panthera leo*. *Animal Behaviour*, 47(2), 379–387.
- McCrink, K., & Wynn, K. (2004). Large-number addition and subtraction by 9-month-old infants. *Psychological Science*, 15, 776–781.
- McCrink, K., & Wynn, K. (2009). Operational momentum in large-number addition and subtraction by 9-month-olds. *Journal of Experimental Child Psychology*, 103, 400–408.
- McCrink, K., Dehaene, S., & Dehaene-Lambertz, G. (2007). Moving along the number line: Operational momentum in nonsymbolic arithmetic. *Perception and Psychophysics*, 69, 1324–1333.
- Mechner, F. (1958). Probability relations within response sequences under ratio reinforcement. *Journal of the Experimental Analysis of Behavior*, 1, 109–121.
- Mechner, F., & Guevrekian, L. (1962). Effects of deprivation upon counting and timing in rats. *Journal of the Experimental Analysis of Behavior*, 5, 463–466.
- Mix, K., Levine, S. C., & Huttenlocher, J. (2003). *Quantitative development in infancy and early childhood*. Oxford, England: Oxford University Press.
- Mix, K., Huttenlocher, J., & Levine, S. (2002). Multiple cues for quantification in infancy: Is number one of them? *Psychological Bulletin*, 128, 278–294.
- Mix, K., Levine, S., & Huttenlocher, J. (1997). Numerical abstraction in infants: Another look. *Developmental Psychology*, 33, 423–428.
- Moore, D., Benenson, J., Reznick, J. S., Peterson, M., & Kagan, J. (1987). Effect of auditory numerical information on infants' looking behavior: Contradictory evidence. *Developmental Psychology*, 23, 665–670.
- Moyer, R. S., & Landauer, T. K. (1967). Time required for judgments of numerical inequality. *Nature*, 215, 1519–1520.
- Naccache, L., & Dehaene, S. (2001). The priming method: Imaging unconscious repetition priming reveals an abstract representation of number in the parietal lobes. *Cerebral Cortex*, 11, 966–974.
- Newcombe, N. (2002). The nativist-empiricist controversy in the context of recent research on spatial and quantitative development. *Psychological Science*, 13, 395–401.
- Nieder, A. (2005). Counting on neurons: The neurobiology of numerical competence. *Nature Reviews Neuroscience*, 6, 177–190.
- Nieder, A., Freedman, D. J., & Miller, E. K. (2002). Representation of the quantity of visual items in the primate prefrontal cortex. *Science*, 297, 1708–1711.
- Nieder, A., & Merten, K. (2007). A labeled-line code for small and large numerosities in the monkey prefrontal cortex. *Journal of Neuroscience*, 27, 5986–5993.
- Nieder, A., & Miller, E. K. (2003). Coding of cognitive magnitude: Compressed scaling of numerical information in the primate prefrontal cortex. *Neuron*, 37, 149–157.
- Nieder, A., & Miller, E. K. (2004). A parieto-frontal network for visual numerical information in the monkey. *Proceedings of the National Academy of Sciences USA*, 101, 7457–7462.
- Opfer, J. E., & Devries, J. (2008). Representational change and magnitude estimation: Why young children can make more accurate salary comparisons than adults. *Cognition*, 108, 843–849.
- Opfer, J. E., & Siegler, R. (2007). Representational change and children's numerical estimation. *Cognitive Psychology*, 55, 169–195.
- Opfer, J. E., & Thompson, C. (2008). The trouble with transfer: Insights from microgenetic changes in the representation of numerical magnitude. *Child Development*, 79, 788–804.
- Opfer, J. E., Thompson, C., & Furlong, E. (2010). Early development of spatial-numeric associations: Evidence from spatial and quantitative performance of preschoolers. *Developmental Science*, 13, 761–771.
- Oyama, T., Kikuchi, T., & Ichihara, S. (1981). Span of attention, backward masking, and reaction time. *Perception and Psychophysics*, 29, 106–112.
- Piazza, M., Facoetti, A., Trussardi, A., Berteletti, I., Conte, S., Lucangeli, D., ... Zorzi, M. (2010, Mar 31). Developmental trajectory of number acuity reveals a severe impairment in developmental dyscalculia. *Cognition*, 116, 1–9.
- Piazza, M., Izard, V., Pinel, P., Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, 44(3), 547–555.
- Piazza, M., Mechelli, A., Price, C. J., & Butterworth, B. (2006). Exact and approximate judgements of visual and auditory numerosity: An fMRI study. *Brain Research*, 1106, 177–188.
- Piazza, M., Pinel, P., Le Bihan, D., & Dehaene, S. (2007). A magnitude code common to numerosities and number symbols in human intraparietal cortex. *Neuron*, 53, 293–305.
- Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian indigenous group. *Science*, 306, 499–503.
- Platt, J. R., & Johnson, D. M. (1971). Localization of position within a homogenous behavior chain: Effects of error contingencies. *Learning and Motivation*, 2, 386–414.
- Quintana, J., Fuster, J. M., & Yajeya, J. (1989). Effects of cooling parietal cortex on prefrontal units in delay tasks. *Brain Research*, 503, 100–110.
- Ramani, G., & Siegler, R. (2008). Promoting broad and stable improvements in low-income children's numerical knowledge through playing number board games. *Child Development*, 79(2), 375–394.
- Räsänen, P., Salminen, J., Wilson, A. J., Aunio, P., & Dehaene, S. (2009). Computer-assisted intervention for children with low numeracy skills. *Cognitive Development*, 24, 450–472.
- Ratcliff, R., Love, J., Thompson, C. A., & Opfer, J. E. (in press). Children are not like older adults: A diffusion model of developmental changes in speeded responses. *Child Development*.
- Rivera, S. M., Reiss, A. L., Eckert, M. A., & Menon, V. (2005). Developmental changes in mental arithmetic: Evidence for increased functional specialization in the left inferior parietal cortex. *Cerebral Cortex*, 15, 1779–1790.
- Sarnecka, B. W., & Carey, S. (2008). How counting represents number: What children must learn and when they learn it. *Cognition*, 108, 662–674.
- Scholl, B. J., & Leslie, A. M. (1999). Explaining the infant's object concept: Beyond the perception/cognition dichotomy. In E. Lepore & Z. Pylyshyn (Eds.), *What is cognitive science?* (pp. 26–73). Oxford, England: Blackwell.

- Sekuler, R., & Mierkiewicz, D. (1977). Children's judgments of numerical inequality. *Child Development*, 48, 630–633.
- Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child Development*, 75, 428–444.
- Siegler, R. S., & Jenkins, E. (1989). *How children discover new strategies*. Hillsdale, NJ: Erlbaum.
- Siegler, R. S., & McGilly, K. (1989). Strategy choices in children's time-telling. In I. Levin & D. Zakay (Eds.), *Time and human cognition: A life span perspective* (pp. 185–218). Amsterdam, Netherlands: Elsevier Science.
- Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science*, 14, 237–243.
- Siegler, R. S., & Ramani, G. B. (2008). Playing linear numerical board games promotes low-income children's numerical development. *Developmental Science*, 11, 655–661.
- Siegler, R. S., & Ramani, G. B. (2009). Playing linear number board games—but not circular ones—improves low-income preschoolers' numerical understanding. *Journal of Educational Psychology*, 101, 545–560.
- Siegler, R. S., & Robinson, M. (1982). The development of numerical understandings. In H. W. Reese & L. P. Lipsitt (Eds.), *Advances in child development and behavior* (Vol. 16, pp. 241–312). New York: Academic Press.
- Siegler, R. S., & Stern, E. (1998). Conscious and unconscious strategy discoveries: A microgenetic analysis. *Journal of Experimental Psychology: General*, 127, 377–397.
- Siegler, R. S., Thompson, C. A., & Opfer, J. E. (2009). The logarithmic-to-linear shift: One learning sequence, many tasks, many time scales. *Mind, Brain, and Education*, 3, 143–150.
- Simon, T. (1997). Reconceptualizing the origins of number knowledge: A "non-numerical" account. *Cognitive Development*, 12, 349–372.
- Simon, T., Hespos, S., & Rochat, P. (1995). Do infants understand simple arithmetic? A replication of Wynn (1992). *Cognitive Development*, 10, 253–269.
- Starkey, P., & Cooper, R. G. (1980). Perception of numbers by human infants. *Science*, 210, 1033–1035.
- Starkey, P., Spelke, E. S., & Gelman, R. (1983). Detection of intermodal numerical correspondences by human infants. *Science*, 222, 179–181.
- Strauss, M., & Curtis, L. (1981). Infant perception of numerosity. *Child Development*, 52, 1146–1152.
- Thompson, C. A., & Opfer, J. E. (2008). Costs and benefits of representational change: Effects of context on age and sex differences in symbolic magnitude estimation. *Journal of Experimental Child Psychology*, 101, 20–51.
- Thompson, C. A., & Opfer, J. E. (2010). How 15 hundred is like 15 cherries: Effect of progressive alignment on representational changes in numerical cognition. *Child Development*, 81, 1768–1786.
- Thompson, C. A., & Siegler, R. S. (2010). Linear numerical-magnitude representations aid children's memory for numbers. *Psychological Science*, 21, 1274–1281.
- Trick, L., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? A limited capacity preattentive stage in vision. *Psychological Review*, 101, 80–102.
- Tudusciuc, O., & Nieder, A. (2009). Contributions of primate prefrontal and posterior parietal cortices to length and numerosity representation. *Journal of Neurophysiology*, 101, 2984–2994.
- Uller, C., Carey, S., Huntley-Fenner, G., & Klatt, L. (1999). What representations might underlie infant numerical knowledge? *Cognitive Development*, 14, 1–36.
- vanMarle, K., & Wynn, K. (2009). Infants' auditory enumeration: Evidence for analog magnitudes in the small number range. *Cognition*, 111, 302–316.
- vanMarle, K., & Wynn, K. (2006). Six-month-old infants use analog magnitudes to represent duration. *Developmental Science*, 9, 41–49.
- Wood, J. N., & Spelke, E. S. (2005). Infants' enumeration of actions: Numerical discrimination and its signature limits. *Developmental Science*, 8, 173–181.
- Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, 358, 749–750.
- Wynn, K. (1996). Infants' individuation and enumeration of actions. *Psychological Science*, 7, 164–169.
- Wynn, K., Bloom, P., & Chiang, W. C. (2002). Enumeration of collective entities by 5-month-old infants. *Cognition*, 83, B55–B62.
- Xu, F. (2003). Numerosity discrimination in infants: Evidence for two systems of representations. *Cognition*, 89, B15–B25.
- Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, 74, B1–B11.
- Xu, F., Spelke, E., & Goddard, S. (2005). Number sense in infants. *Developmental Science*, 8, 88–101.

Visuospatial Thinking

Mary Hegarty and Andrew T. Stull

Abstract

Visuospatial thinking includes thinking about space at the smaller scale of objects and at the larger scale of environments. It also includes situations in which we use visuospatial representations to think about nonspatial entities. At the scale of objects, this chapter reviews the types of representations and processes that underlie object recognition and categorization, the nature of visuospatial mental images of objects, and how these are processed in interactions with objects and in reasoning and problem solving. At the scale of environments, we examine the nature of environmental spatial representations and the processes that operate on these to keep us oriented in space, to reorient ourselves when we are lost, to learn the layout of new environments, and to plan routes through familiar environments. Finally, the chapter reviews how spatial representations are used metaphorically, to think about nonspatial entities, in language, reasoning, and graphics.

Key Words: imagery, navigation, spatial updating, cognitive maps, spatial metaphors, diagrams, graphs

Introduction

When we find our way back to our parked car at the end of the day, think about how to rearrange the furniture in our living room, or read a graph, we are engaging in visuospatial thinking. Visuospatial thinking is central to many scientific domains and professions. For example, chemists think spatially when they develop models of the structure of molecules to understand their reactive properties, geologists think spatially when they reason about the physical processes that form mountains and canyons, and architects think spatially when they design a new house. Visuospatial thinking involves thinking about the shapes and arrangements of objects in space and about spatial processes, such as movement and deformation. It also involves maintaining a representation of our location and orientation with respect to the larger environment, updating how that changes as we move through space, and planning routes. Finally, it includes thinking with

spatial representations of nonspatial entities, such as spatial metaphors, diagrams, and graphs.

Theories of cognition recognize the centrality of visuospatial thinking. For example, in working memory theories, one of the most fundamental distinctions is between visuospatial versus verbal working subsystems (Baddeley & Lieberman, 1980; Shah & Miyake, 1996). Similarly, in research on intelligence, spatial visualization was identified as one of the seven primary mental abilities (Thurstone, 1938) and is one of Gardner's (2004) multiple intelligences. Spatial thinking is fundamental for survival and is not uniquely human (see Penn & Povinelli, Chapter 27). Animals keep a record of their position and orientation in the environment, find their way back to their nests at the end of the day, and often demonstrate feats of spatial intelligence that exceed those of humans (Muller & Wehner, 1988). Pigeons can recognize objects from novel perspectives (Friedman, Spetch, & Ferrey, 2005). Crows

use tools and cache food throughout their environments, locating it several months later (Emery & Clayton, 2004).

The title of this chapter is visuospatial thinking, reflecting the fact that vision is the primary sense by which we sense spatial properties of the world, and research on visuospatial thinking has traditionally focused on thinking with visual perceptions and images. However, not all visual properties are spatial properties. For example, the shape and location of an object are spatial properties, but its color and brightness are not. Furthermore, spatial properties are not just sensed by vision. We can sense shapes of objects by haptic exploration, locate people by hearing their voices, and update our location in the environment by integrating body-based senses, and there is growing evidence that spatial representations are often multimodal or amodal (e.g., Avraamides, Loomis, Klatzky, & Golledge, 2004). While focusing on visuospatial thinking, our chapter will also address how other perceptual modalities and motor processes allow humans to think about space.

We interact with space in different ways at different scales. Montello (1993) distinguishes between *figural* space, which is small in scale relative to the body and external to the individual (e.g., the space of objects); *vista* space, which is projectively as large or larger than the body and contains the individual but can be visually apprehended from a single place (the space of rooms or scenes); and *environmental* space, which is apprehended over time by moving through the environment (the space of buildings, towns, or cities). Spelke and Kinzler (2007) distinguish between four core knowledge systems, including an object system that represents the cohesion of object properties and mechanical interactions between objects and a spatial system for representing the geometry of the environment (their other two proposed core systems represent agents and their goal-directed actions, and number, respectively). Previc (1998) proposed four realms for spatial behaviors: peripersonal (near-body space), focal extrapersonal (the space of visual search and object recognition), action extrapersonal (orienting in topographically defined space), and ambient extrapersonal (orienting in earth fixed space).

In this chapter, we distinguish between spatial thinking at two broad scales of space: (1) small-scale or object-based space, which includes activities such as imagining object transformations and planning interactions with objects, and (2) large-scale or environmental space, which includes activities

such as learning the layout of a new environment, or planning a route. Spatial thinking at both scales involves perceiving spatial properties, maintaining spatial representations in working memory, and transforming those representations. However, we interact differently with the world at these scales. We manipulate objects within reach of our bodies (peripersonal space) but not objects outside of that space. These scales typically involve different frames of reference; in our interactions with objects, we generally think of ourselves as stationary while the objects move, whereas in our interactions with environments we think of ourselves as moving and the environment as stationary. Moreover, there is evidence for dissociations between spatial thinking at these two scales of space, in both individual differences (Hegarty, Montello, Richardson, Ishikawa, & Lovelace, 2006) and in terms of neural systems. Object-based transformations such as mental rotation depend primarily on areas in the superior parietal cortex (Zacks, 2008), whereas environmental spatial tasks depend on a network of brain regions in the medial temporal cortex, parietal cortex, and retrosplenial cortex (Byrne, Becker & Burgess, 2007; Epstein, 2008).

Another distinction that we make in this chapter is between thinking about space and using space to think. Visuospatial thinking includes situations in which we use spatial representations to think about other entities, both abstract and concrete. For example, we follow the path of life, feel “down” when we are sad, and climb the corporate ladder (Lakoff & Johnson, 1980). We also use spatial representations to reason, for example, when we represent premises in a reasoning problem as Euler circles (Stenning & Oberlander, 1995) or use diagrams, maps, and graphs, which enable us to “use vision to think” (Card, Mackinlay, & Schneiderman, 1999).

This chapter is therefore divided into three main sections. The first two address thinking about space, at the smaller scale of objects and the larger scale of environments, respectively. The third section is about using space to think. They correspond loosely to what a recent National Research Council report (2006) refers to as thinking about space, thinking in space, and thinking with space.

Thinking About Space: The Space of Objects

In considering visuospatial thinking at the scale of objects, we first briefly review the types of representations and processes that underlie object recognition

and categorization. We then examine the nature of visuospatial mental images and how these are used when we interact with objects and in more complex processes of reasoning and problem solving.

Recognizing and Categorizing Objects

Recognizing and categorizing objects in the external world are not trivial tasks, which may not seem obvious given the speed and apparent ease with which we accomplish them (Thorpe, Fize, & Marlot, 1996). Recognition of an object involves edge detection, figure-ground perception, and binding of features such as color, contrast, and shading into a common form that is understood to be a three-dimensional object. The complexity of this process is compounded when considering that the image of an object on the retina is not likely to be identical on any two occasions because of differences in viewpoint, distance, and so on (Hayward, 2003; Lawson, 1999). Categorization adds further complexity, because there is great variability of size, shape, and texture within members of a single category. For example, our ability to categorize armchairs, office chairs, and kitchen chairs as chairs requires considering some features of the object while ignoring others.

MODELS OF OBJECT RECOGNITION

Visuospatial cognitive processes are central to the long-standing debate over the mechanisms by which objects are recognized (Biederman, 1987; Hayward, 2003; Tarr & Pinker, 1989). One class of object recognition model, structural description models (also known as viewpoint invariant models) (Hayward, 2003), posits that viewed objects are recognized by comparing them to schematic viewpoint invariant representations based on either their principal axes (Marr & Nishihara, 1978) or the spatial relationship between component parts (i.e., volumetric primitives called geons) (Biederman, 1987). These models predict that time and accuracy to recognize a previously viewed object will not be affected by viewpoint differences as long as the invariant features and their spatial relationships are visible.

A second class of model, referred to as view-based or view-dependent models, posits that recognition involves the comparison of a two-dimensional projection or image encoded from the observed viewpoint with a previously encoded two-dimensional image stored in memory and retrieved based on common features (Tarr & Pinker, 1989). This class of models specifies that multiple views of the object,

influenced by the observer's experience with that object, are stored in memory. To recognize an object from an unfamiliar view, an analog normalization process is proposed to transform the stored image to match the percept (Humphreys & Riddoch, 1984; Lawson, 1999; Lawson & Humphreys, 1999). These models predict that recognition time and accuracy will be proportional to the angular difference between the stored and viewed image.

Given that there is evidence for both viewpoint-dependent and viewpoint-independent recognition performance under different circumstances, object-recognition researchers have begun to develop hybrid models (e.g., Foster & Gilson, 2002; Hayward, 2003; Hummel, 2003; Vanrie, Willems, & Wagemans, 2001). For example, Graf and colleagues (Graf, 2006, 2010; Graf, Kaping, & Bülthoff, 2005) have proposed a model in which the stored representation is schematic, as in structural description models, but viewpoint dependent, as in view-based models; and it is this schematic representation, rather than a two-dimensional visual image, that is transformed by an analog transformation when the view to be recognized is different from the stored view. The primary support for this model is the *congruency effect*, which is that sequentially presented objects are faster to recognize if their main axes are aligned, even for objects of dissimilar categories and form (Gauthier & Tarr, 1997; Jolicoeur, 1990, 1992).

Extending this idea, transformation of a perceptual coordinate system may also account for the similarities between object recognition and categorization. Graf (2010) proposes that recognition involves comparing the stimulus percept with the memory representations after Euclidian transformations (i.e., rotation, size, and displacement), whereas categorization involves additional non-Euclidian transformations that deform or morph the percept of an object to match a stored representation. In this light, the physical space of an object, the perceptual space encoded from the stimulus, and the representational space stored in memory might share a topological structure and object recognition and categorization might be based on common neural processes (Panis, Vangeneugden, & Wagemans, 2008).

Imagining Objects

PERCEPTUAL CHARACTERISTICS OF IMAGES

In contrast with object recognition and categorization, visual imagery refers to the experience of visualizing something that is not physically present, so that

there is no corresponding sensory input to the cognitive system. The experience of having a visual image feels similar to the experience of seeing. For example, most people report that they experience imagery when answering questions such as “how many windows are in your home?” Objective measures also reveal several parallels between imaging and seeing. The time to answer questions about objects in a mental image is correlated with the relative size of those objects, as if one has to “zoom” into the image to see the details of the object’s appearance (Kosslyn, 1980). Time to scan between objects in a mental image is also related to the distance between these objects, just as it takes more time to scan between objects that are farther apart when looking at a real scene (Finke & Pinker, 1983; Kosslyn, Ball, & Reiser, 1978).

Despite these similarities, there are also important differences between mental images and visual percepts. Images are generally less vivid and detailed (Chambers, 1997). Images are also internally organized so that it is easier to “see” certain subcomponents of images than others. For example, it is easier to see a triangle than a parallelogram in an image of the Star of David (Reed, 1974). Furthermore, people cannot always reinterpret patterns in images as well as they can reinterpret them in pictures (Chambers & Reisberg, 1985, 1992), although some ambiguous figures can be reinterpreted (Finke, Pinker, & Farah, 1989) and reinterpretation can be an important cognitive process in creative thinking (Finke, 1989).

THE IMAGERY DEBATE

While the phenomenology of visual imagery is not at issue, there has been much debate in psychology about the nature of the internal representations that underlie the experience of mental imagery (for recent reviews, see Kosslyn, Thompson, & Ganis, 2006; Pylyshyn, 2003). According to Kosslyn et al. (2006), mental images reflect a distinct type of internal representation, a *depictive* representation, in which the representation resembles the object represented, parts of the representation depict parts of the object, the shapes of these parts and their spatial relationships correspond to the shapes and spatial relationships of the represented objects, and so on. In contrast, Pylyshyn (2003) proposes that what distinguishes images is their content, not their format. According to his view, there is no reason to postulate a separate form of representation underlying imagery and when we experience mental imagery, we are simulating what something would look like, based on *tacit knowledge*, or what we have come

to know about the appearances of objects from our visual experiences, but our conscious experience of an image may arise from representations that are not themselves depictive.

NEURAL BASIS OF IMAGERY

In recent years, the imagery debate has been fueled by findings that both primary and secondary visual cortex are activated during many imagery tasks (Kosslyn & Thompson, 2003). These cortical areas are topographically mapped such that specific neurons represent specific areas of space. Thus, representations in these cortical areas are depictive, in the sense that parts of the cortical representation correspond to parts of the represented object. The patterns of activation in these areas during imagery have been shown to mimic perception. For example, when people view objects that subtend smaller visual angles, activation is more posterior in primary visual cortex than for larger percepts. The same is true when people construct smaller versus larger images (Kosslyn, Alpert, Thompson, & Maljkovic, 1993). Similarly, patients with brain lesions that affect visual perception are also impaired in visual imagery (Farah, 1988), and hemispatial neglect patients ignore the same areas of space in perception and imagery (Bisiach & Luzatti, 1978).

Although many imagery tasks activate early visual cortex, not all do so. A recent review of neuroimaging studies (Kosslyn & Thompson, 2003) revealed three characteristics of studies in which activation was found in these areas: (1) the task involved inspecting high-resolution details of images, (2) the task required the visualization of shapes, and (3) the measurement technique was particularly sensitive. These results highlight that not all visual imagery tasks have the same demands. Some imagery tasks, such as answering questions about the shapes of letters or the colors of objects require people to inspect details of the appearance of objects and are accompanied by activation of primary visual cortex. Other imagery tasks, such as mental rotation or making judgments about the spatial locations of objects activate parietal rather than visual cortex and can be performed as well by congenitally blind as by sighted people, suggesting that they may rely on multimodal or amodal representations rather than specifically “visual” representations.

Interestingly, individual-difference studies also suggest dissociations between ability to perform these *visual* and *spatial* tasks (Kozhevnikov, Blazhenkova, & Becker, 2010). There are large individual differences

in vividness of visual imagery, first studied by Galton (1883) when he asked people to imagine their breakfast table. There are also large individual differences in ability to imagine spatial transformations of images, as measured by spatial ability tests (Hegarty & Waller, 2005). But these individual differences dimensions are uncorrelated (Kosslyn, Brunn, Cave, & Wallach, 1984). The dissociation between visual and spatial aspects of mental images is thought to reflect a division of labor in the visual system between a ventral, *object* properties pathway (otherwise known as the “what” subsystem) that projects from the occipital to the inferior temporal lobe, and a dorsal, *spatial* properties pathway, otherwise known as the “where” subsystem (Ungerleider & Mishkin, 1982) or the “how” subsystem (Goodale & Milner, 1992) that projects from the occipital lobe to the posterior parietal lobe.

TRANSFORMATIONS OF IMAGES

Thinking and reasoning about objects often involves mentally transforming images. While there are many ways in which we can transform images, by adding or subtracting parts, size scaling of parts of the image, and so on, most research on this topic has focused on mental rotation of images, so we will consider this transformation in detail. When people are asked to judge whether the two objects in Figure 31.1 depict the same object or are mirror images of each other, the time taken to answer this question is linearly related to the angular difference in orientation between the two objects, suggesting that people mentally rotate a visual image of one of the forms into congruence with the other (Shepard & Metzler, 1971). This rotation process was proposed to be analog such that when objects are mentally rotated, they are imagined at all intermediate orientations between the start and final orientation (Cooper, 1976; Cooper & Shepard, 1984). For example, if people are interrupted during the mental rotation process, and their rate of mental rotation is known, they show no additional time to respond to a stimulus that matches the amount of rotation that should have occurred at the time of the interruption. Early studies also suggested that mental rotation is holistic. Thus, differences in complexity of two-dimensional shapes did not affect rotation times (Cooper & Podgorny, 1976). However, later studies suggested that the lack of a complexity effect might have occurred because the task could be accomplished by rotating only a subset of the stimulus information (Folk & Luce, 1987; Yuille & Steiger, 1982) and that rotation is piecemeal when the full

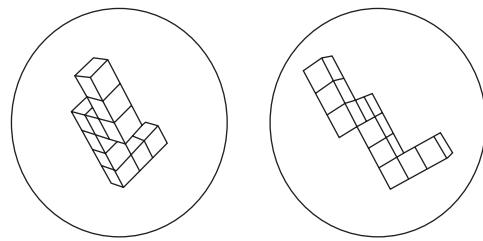


Fig. 31.1 Example of a trial in a mental rotation study.

complexity of the stimulus needs to be considered or the stimuli are less novel (Bethell-Fox & Shepard, 1988). Eye fixation studies also suggested a piecemeal strategy (Just & Carpenter, 1976).

Mental rotation strategies and performance are affected by individual differences in spatial ability. For example, in a cube comparison task, Just and Carpenter (1985) found that when the most direct rotation is around an axis that is nonorthogonal to the axes of the object, high-spatial individuals rotate around this axis, whereas low-spatial individuals perform a sequence of rotations around orthogonal axes. Other research has shown that rotations around axes that are orthogonal to axes of the object and environment are generally easier than rotations around nonorthogonal axes (Pani, 1993; Pani, William, & Shippey, 1995). In summary, the literature on mental rotation suggests that while mental rotation involves analog imagery, other factors such as familiarity with the objects, experience with the task, and spatial ability enable more analytic strategies that can be used in conjunction with imagery to simplify the imagery process.

IMAGERY OF HANDS AND INTERACTIONS WITH OBJECTS

Our hands are our primary tool for exploring and modifying objects and, therefore, they play a special role in visuospatial cognition (Halligan, Fink, Marshall, & Vállar, 2003). Time to rotate an image of a hand is not linearly related to angle of displacement (cf. Shepard & Metzler, 1971). Rather, the constraints of real hand movement affect the response time and more time is required to imagine rotations of hands through orientations that are biomechanically awkward (Parsons, 1987a, 1987b, 1994; Sekiyama, 1982). Observing hand rotation also facilitates motor imagery (Conson, Sarà, Pistoia, & Trojano, 2009), and this facilitation is specific to the stimulus (hands) and motion (rotation).

Other research suggests that mental rotation shares representations and processes with planning

and executing physical interactions with objects (Wohlschläger, 2001). Manual rotation facilitates mental rotation of objects if the directions of the two rotations are congruent and interferes when the directions of the rotations are incongruent (Wexler, Kosslyn, & Berthoz, 1998; Wohlschläger & Wohlschläger, 1998; but see Sack, Lindner, & Linden, 2007). Furthermore, actions when using a tool facilitate mental rotation even when the action is not a rotation. For example, pulling the string wrapped around a spool (a linear translation action) facilitates mental rotation of an object lying on the spool (Schwartz & Holton, 2000). Finally, the motor cortex is engaged when participants imagine rotating an object with their hand, but not when they do not imagine their hand as the agent (Kosslyn, Thompson, Wraga, & Alpert, 2001).

More generally, the locations of hands in space affect how we allocate attention, whether the hands belong to the viewer, another human, or are disembodied hand-shaped objects (Reed, Grubb, & Steele, 2006). This biased attention is more evident in the graspable region (in front of the hand) than other regions equally close to the hands or tools (e.g., back of the hand) (Reed, Betz, Garza, & Roberts, 2009). In fact, the actions that we can perform within reach of our hands help to divide the space that surrounds our body. The space that is in reach (peripersonal space) is dissociable and handled by different neural systems than the space that is outside our reach (extrapersonal space), but peripersonal space is based on the potential for action rather than being a fixed size. For example, it is smaller for amputees with restricted reach (Makin, Wilf, Schwartz, & Zohary, 2009) and is expanded to include the space surrounding tools that extend actionable space (Berti & Frassinetti, 2000; Carlson, Alvarez, Wu, & Verstraten, 2010; Halligan et al., 2003; Macaluso & Maravita, 2010; Maravita & Iriki, 2004). In summary, imagery, motor planning, and attention are interacting cognitive processes unified because of the presence of our hand in the actionable space around our body.

Imagery in Complex Spatial Tasks

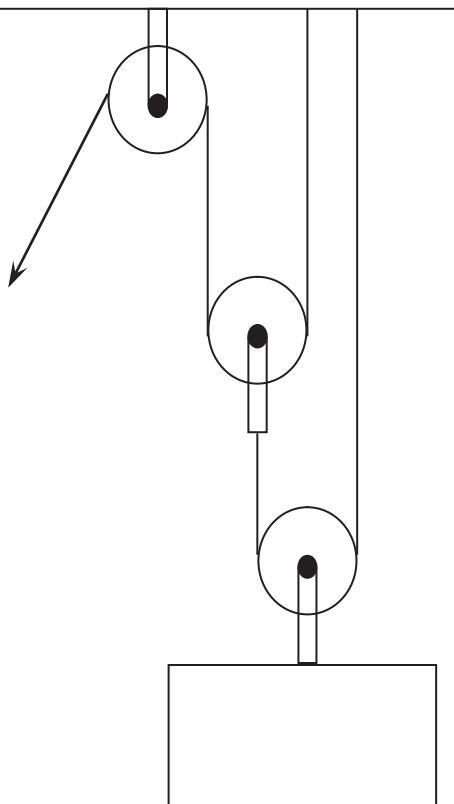
Although a great deal of research has examined how we imagine basic spatial transformations, and mental rotation in particular, the visuospatial tasks that we carry out in everyday life involve imagining more complex spatial transformations. They often require imagining a series of spatial transformations, as when we assemble a piece of furniture

(Tversky et al., 2007) or rearrange the furniture in our living room. Similarly, in diagnosing faults in your car engine, mechanics have to imagine different motions (rotations, translations, ratcheting, etc.) of many components, and scientists have to imagine complex spatial transformations when they reason about flying balls and spinning gyroscopes, or justify the relationship between molecular structure and chemical reactivity (Stieff, 2007). These do not just involve mentally rotating one rigid object (as in measures of mental rotation). In these situations, it appears that analog mental imagery transformations are augmented by analytic processes, including task decomposition and rule-based reasoning.

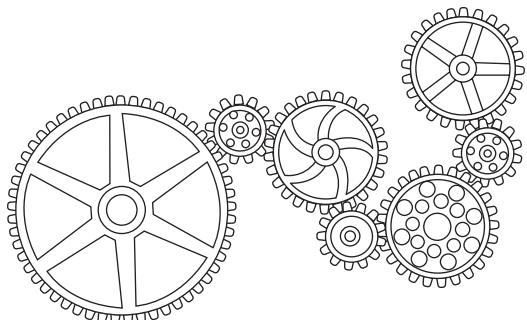
TASK DECOMPOSITION

One way in which people augment analog imagery processes in visuospatial thinking is by task decomposition. They mentally simulate complex motions piecemeal rather than holistically. For example, if you pulled on the rope of the pulley system illustrated in Figure 31.2a, all of its parts would move together. However, Hegarty (1992) showed that people infer the motion of a mechanical system like this by decomposing the task into a sequence of relatively simple interactions (e.g., how the motion of a rope causes a pulley to rotate) and inferring the motion of one component at a time. The time taken to verify how a component moved was thus proportional to its position in this causal chain.

At the same time, several lines of evidence suggest that people often use imagery transformation processes in mechanical reasoning when they imagine each “link” in the causal chain, rather than merely applying verbally encoded rules. Visuospatial working memory loads interfere more with mechanical reasoning tasks like the pulley problem than do verbal working memory loads (Sims & Hegarty, 1997). Similarly, mechanical reasoning interferes more with visuospatial than with verbal memory loads, suggesting that mechanical reasoning depends on representations in visuospatial working memory (cf. Logie, 1995). When asked to “think aloud” while they infer how parts of a machine work, people tend to express their thoughts in gestures rather than in words (Hegarty, Mayer, Kriz, & Keehner, 2005; Schwartz & Black, 1996; see also Goldin-Meadow & Cook, Chapter 32), and asking people to trace an irrelevant spatial pattern while reasoning about mechanical systems impairs their reasoning (Hegarty et al., 2005). In summary, mechanical



a. Pulley System



b. Gear System

Fig. 31.2 Examples of pulley and gear systems. In mechanical reasoning problems, one might be asked which direction the lower pulley moves (clockwise or counterclockwise) when the rope is pulled or one might be told that the large gear on the left is moving clockwise and asked to predict the direction of the right uppermost gear.

reasoning involves first decomposing the task into one of inferring the motion of individual components and then using mental image transformations to simulate the motion of each component in order of the causal chain of events (Hegarty 1992, 2004).

AUGMENTATION BY RULE-BASED REASONING

Another way in which imagery-based processing is augmented by more analytic thinking is that it sometimes leads to the induction of regularities, so that rule-based reasoning takes over. For example, take the gear problem in Figure 31.2b. When Schwartz and Black (1996) asked people to solve problems like this, participants' gestures indicated that people initially mentally simulated the motion of the individual gears, but on the basis of these simulations, they discovered the simple rule that any two interlocking gears move in opposite directions. Participants then switched to a rule-based strategy, but they reverted to the mental simulation strategy again when given a novel type of gear problem. Replacing simulation-based reasoning with rule-based reasoning was also observed in problem solving with diagrams in organic chemistry (Stieff, 2007; Stieff & Raje, 2010) and in psychometric tests of spatial ability (e.g., Hegarty, 2010; Lohman, 1988). In these situations, it has been proposed that imagery strategies are domain-general problem-solving heuristics used by novices, whereas rule-based analytic strategies are discovered during task performance or instruction and are more evident in expert problem solving (Schwartz & Black, 1996; Stieff, 2007; Stieff, Hegarty, & Dixon, 2010). However, experts also use mental simulation in novel situations in which their rules are inadequate (e.g., too narrow for the situation at hand), and in some domains such as algebra, the development of expertise is characterized by a shift from more abstract rule-based thinking to simulation-based thinking (e.g., Goldstone, Landy, & Son, 2010), so the development of expertise might be best seen as a process of developing the least effortful strategies, whether imagistic or analytic, for different tasks.

REVEALING IMPLICIT KNOWLEDGE

Research on mechanical reasoning and problem solving has suggested that an important function of imagery may be to “reveal” or make available knowledge about properties such as the shapes of objects or motion constraints that is otherwise tacit or implicit. For example, consider the problem in

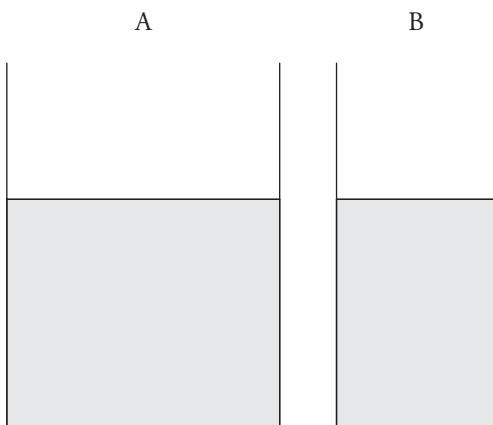


Fig. 31.3 Water Pouring Problem (Schwartz & Black, 1999). The water level in the two glasses is the same. If you start to rotate the two glasses at the same rate (degrees of rotation per second), in which glass will the water start pouring out first? A or B?

Figure 31.3, which people had to judge the angle at which water would pour out of two glasses, a fat glass and a thin glass. People perform very poorly on this task when they answer on the basis of explicit knowledge. However, when they close their eyes and rotate an imaginary glass to indicate the answer, they can almost always correctly judge that water will start pouring out of the fat glass first, and there is no systematic relationship between their answers when they are and are not instructed to use imagery (Schwartz & Black, 1999). Visualizing the situation reveals knowledge of which the person was otherwise unaware. The same process seems to occur when we take a mental walk to “discover” how many windows are in our house. Furthermore, when imagery is externalized as gesture, it can also reveal knowledge to others, and observing one’s own gestures might also reveal implicit knowledge to oneself, and this is one possible way in which gesture enhances thinking (see Goldin-Meadow & Cook, Chapter 32). In these situations, it seems that the knowledge was there to begin with, but because it was initially linked to our perceptual experience, it may take an act of mental imagery or simulated physical action to make it available to our thought processes.

In summary, simulating spatial transformations using visual imagery is an important strategy in tasks such as mental rotation, mechanical reasoning, and chemistry problem solving, but in each of these cases, more analytic strategies are also used. These can involve decomposing the problem, such that less information needs to be visualized at a time, abstracting nonspatial information, and the

application of rules to generate an answer or eliminate answer choices. One possibility is that mental imagery transformations are effortful processes, and the best spatial thinkers are those who augment visualization with more analytic strategies, using these analytic strategies when they can, so that they rely less on effortful imagery processes.

Thinking About Space: The Space of Environments

The ability to navigate in the world is a fundamental cognitive function. Navigation is necessary for finding our way in familiar environments, learning the layouts of new places, and planning routes to distant locations. Navigation can be based on external spatial representations such as maps and on internal spatial representations derived from sensory experience. It is a multisensory process in which information needs to be integrated and manipulated over time and space. Vision provides us with direct information about spatial attributes of the environment, such as its general shape, and the objects therein. However, in most natural environments, not all relevant features can be seen from a single vantage point. Therefore, keeping track of one’s location and orientation is crucial if we are to integrate all relevant features into a comprehensive representation, so that navigation also depends on perceiving self-motion.

Wolbers and Hegarty (2010) provide an overview of the sensory cues, perceptual and cognitive processes, and spatial representations involved in most forms of human and animal navigation. The sensory cues include cues in the environment, such as landmarks (e.g., a distinctive building), global orientation cues (e.g., the sun), the geometric structure of the environment (e.g., the network of streets in a city), and symbolic representations (e.g., maps or verbal directions). They also include self-motion perception based on vestibular and proprioceptive cues, and optic flow. The perceptual and cognitive processes that act on these sensory cues include space perception, self-motion perception, computing distances and directions to unseen places, and imagining shifts in perspective or reference frame. Wayfinding and navigation can also depend on executive processes such as selection and maintenance of navigational goals or choice of a navigation strategy (e.g., depending on familiar routes versus global orientation cues). Finally, navigation depends on a variety of spatial representations, including online or short-term representations that are available

when one is navigating in an environment and more enduring offline or long-term memory representations. We first discuss the nature of these representations and then review the perceptual and cognitive processes that operate on them to keep us oriented in space, to reorient when we are disoriented or lost, to learn the layout of new environments, and to plan routes through familiar environments.

Spatial Representations in Orientation and Navigation

Orientation and navigation in the environment rely on two types of spatial representations, egocentric and allocentric representations. Egocentric representations encode the distances and directions to objects and features in the environment with respect to the self, are relatively precise, and depend on areas of the parietal cortex (Byrne et al., 2007). They are primarily online spatial representations that are active in working memory when we are actually in and oriented to an environment.

Allocentric representations encode the locations of objects in the environment with respect to other objects or features of the environment, and they are thought to depend more on areas in the medial temporal lobes (Byrne et al., 2007). Allocentric representations support our long-term representations of the layout of environments that endure when we are no longer in or oriented to the environment and are less precise than online egocentric representations (Waller & Hodgson, 2006). They are often referred to as cognitive maps, so it is important to identify the ways in which they are unlike printed maps. For example, cognitive maps (especially maps of large-scale spaces that cannot be apprehended in a single view) encode locations and distances categorically and hierarchically (Friedman, Brown, & McGaffey, 2002; Hirtle & Jonides, 1985; Stevens & Coupe, 1978), distort distances (Holyoak & Mah, 1982; Sadalla, Burroughs, & Staplin, 1980), and encode the spatial relations between features as more aligned than they are in reality (Tversky, 1981).

Our cognitive maps differ depending on how they are learned. When we learn the layout of an environment from a map, the orientation of the map (usually North is up) defines a preferred orientation of our cognitive map in memory (Levine, Marchon, & Hanley, 1984; Richardson, Montello, & Hegarty, 1999). Preferred orientations are inferred when people are better able to perform pointing tasks (known as judgments of relative directions) from some imagined orientations in the environment than

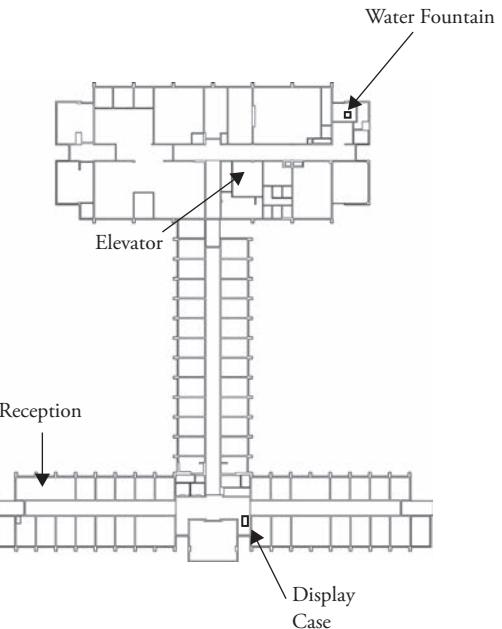


Fig. 31.4 Map of the corridors of a building with North up orientation.

from other imagined orientations. For example, if you learn the layout of a building from the map in Figure 31.4, it is more difficult to point to objects if you imagine yourself at the elevator facing the display case than if you imagine yourself at the display case facing the elevator.

Even when people learn the layout of an environment by direct experience, their enduring representations of the environment can have one or more preferred orientations in memory (e.g., Mou & McNamara, 2002; Shelton & McNamara, 2001). On the basis of extensive studies showing that judgments of relative direction are easier from some imagined orientations in an environment than others, McNamara and colleagues have proposed that object locations in an environment are represented with respect to spatial reference frames with an intrinsic axis. This intrinsic axis can be defined by the shape of the environment (e.g., the axis defined by the longer walls of a rectangular room), a more local reference system (e.g., a mat on the floor, Shelton & McNamara, 2001), or by the configuration of objects themselves, such as an axis of symmetry in the layout of a group of objects (Mou & McNamara, 2002). In the absence of these cues, the orientation from which one has experienced the environment typically defines the axis of the reference frame (Diwadkar & McNamara, 1997). Effects of experience in the environment can also

interact with aspects of the environment (Shelton & McNamara, 1997).

Processes in Orientation and Navigation

REMAINING ORIENTED TO AN ENVIRONMENT

When people are oriented to an environment, their actual facing direction in the environment strongly determines the preferred direction of their online representations. Ability to imagine a perspective other than your current orientation in the environment is effortful and related to the angular disparity between your actual direction and the direction to be imagined (Hintzman, O'Dell, & Arndt, 1981; Rieser, 1989). While perspective taking has been proposed to reflect a mental transformation process analogous to mental rotation (e.g., Easton & Sholl, 1995; Rieser, 1989), an alternative account suggests that the difficulty may reflect a conflict between one's actual directional relations to the objects in the environment and one's relation to those objects from the imagined perspective (e.g., May, 1996; Presson, 1987). Consistent with the interference account, when people are disoriented after viewing an environment (so that they do not know their facing direction), the effects of mismatch between their actual orientation and the orientation to be imaged are reduced (May, 1996; Waller, Montello, Richardson, & Hegarty, 2002). This account may also explain the fact that mental rotation ability and perspective-taking abilities are partially dissociated (Hegarty & Waller, 2004; Kozhevnikov & Hegarty 2001).

As we move about the world, internal and external self-motion cues allow us to keep track of the locations of external features with respect to the self in egocentric coordinates, a process known as spatial updating. These cues include optic flow, changes in the relative locations of objects as we move, as well as body-based self-motion cues (proprioception, motor efference copy, and the vestibular sense). Body-based cues are sufficient for spatial updating over short distances. For example, Rieser (1989) found that pointing from an imagined orientation in an environment was no more difficult than pointing from a studied view, as long as people physically turned to the imagined view, while blindfolded, after seeing the environment from the studied view (see also Presson & Montello, 1994; Rieser, 1989; Simons & Wang, 1998; Waller et al., 2002; Wang & Simons, 1999). In contrast, merely imagining moving from the studied to the imagined view, without actually moving, is not sufficient to

update one's orientation (Klatzky, Loomis, Beall, Chance, & Golledge, 1998). Body-based cues can also be hard to ignore. For example, Farrell and Robertson (1998) had people point to previously seen objects in an environment after being blindfolded and rotated in the environment and found that they had difficulty ignoring their body rotation to point from the original, viewed orientation. They suggested that updating is automatic. However, Waller et al. (2002) found that people could easily follow instructions to either imagine leaving the environment behind them when they rotated or to bring the environment with them, suggesting that updating is not automatic but instead can reflect cognitive awareness of how one's movement affects the task at hand.

While there is good evidence that remaining oriented to an environment involves updating egocentric representations, there has been controversy regarding whether people also update their orientation and location with respect to allocentric representations (Waller & Hodgson, 2006; Wang & Spelke, 2000). In studies on this topic, people first learn an array of objects and then point to the (no longer visible) objects before moving, after rotating in place, or after disorientation. It is argued that if each of the objects is updated with respect to an egocentric frame of reference, the configuration of objects should be distorted after movement, whereas if updating is with respect to an allocentric representation, the configuration of objects should remain intact in memory after movement. Wang and Spelke (2000) found that the configuration of objects (but not aspects of room geometry) was distorted in memory following disorientation, and they concluded that remaining oriented to the environment depends primarily on egocentric representations. Waller and Hodgson (2006) offered an alternative interpretation, that the increased disorientation in memory reflects a switch from more precise online representations to less precise offline representations. They showed that our offline representations of even very familiar environments (our bedrooms) were less precise than online representations of novel environments, and their research also suggested that the switch from online to offline representations could occur even without disorientation (after a rotation of 90 degrees or more). The relative dependence on egocentric versus allocentric representations can also depend on the geometric regularity of object layouts in an environment (Xiao, Mou, & McNamara, 2009).

REORIENTING AFTER DISORIENTATION

In everyday life, we have to reorient to the larger environment when we emerge from a building, such as a department store, through a different door than which we entered. More generally, while we can update our orientation and location in environments over short distances, error accumulates over large distances so that reorientation is often necessary.

There has been much research regarding the cues people use to reorient to an environment after disorientation. In classic research on this topic, Cheng (1986) and Gallistel (1990) found that rats used the general shape of environments (environmental geometry) but not distinctive landmarks to reorient. They placed rats in a rectangular enclosure including distinct odors and visual cues until they found food in one of the corners of the enclosure. Before they had eaten all the food, the rats were removed, disoriented, and placed in an identical enclosure, and it was observed where they searched for food. The surprising result was that the rats looked equally in the correct corner and the diagonally opposite corner (see Fig. 31.5), even though there were salient olfactory and visual cues that identified the correct corner. That is, they searched based on geometry alone and ignored featural cues.

Hermer and Spelke (1994) studied the same paradigm with children and found that young children (younger than 2 years) also ignored highly salient cues and reoriented to environments based on geometry alone. Older children who had acquired spatial language (especially the terms “left” and “right”) successfully combined geometric and featural cues. Moreover, when adults performed a verbal secondary task that prevented them from using language, they also reoriented based on geometry alone (Hermer-Vasquez, Spelke, & Katsnelson, 1999). On the basis of these results, they proposed that humans and animals share a “geometric module” that only processes environmental shape, but that language allows

humans to penetrate this module to combine geometric with featural information. However, recent research has questioned this account. For example, nonhuman species (who obviously do not have language) can use features in some circumstances, younger children do use features in larger environments than those tested by Hermer and Spelke, and nonverbal secondary tasks also impair adults’ performance (see Cheng & Newcombe, 2005; Twyman & Newcombe, 2010 for reviews). An alternative current account of reorientation is that it depends on an adaptive combination of cues (both geometric and featural), based on their salience, reliability, and familiarity (Cheng, Shettleworth, Huttenlocher, & Rieser, 2007), rather than a module that can only take geometry into account.

LEARNING SPATIAL LAYOUT

There has been much interest in how our representations of space develop with increasing experience in that space. One influential theory (Siegel & White, 1975) proposed that we construct qualitatively different types of representations at different stages in our familiarity with an environment; we initially construct landmark representations that allow us to recognize salient landmarks in the environment, later develop route representations that encode the relations between pairs of landmarks that one would encounter as one moves through an environment, and finally develop survey or configurational representations specifying the relations between locations in the environments, independent of the routes between them. Survey representations allow us to take novel shortcuts through the environment and to point directly or estimate straight-line distances to locations in the environment.

While distinguishing between landmark, route, and survey representations of large-scale space has been useful theoretically, this sequential theory by Siegel and White (1975) has been challenged

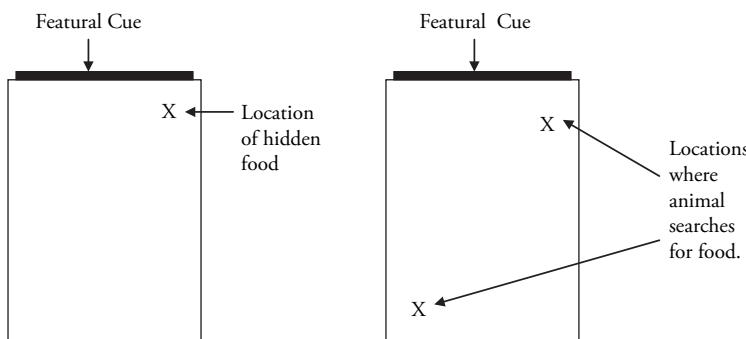


Fig. 31.5 Example of a trial from the experiments of Cheng (1986) showing the location at which food was hidden, and the locations in which the animal searched for food. Although there was a distinct cue that could discriminate the corner in which the food was hidden, the animal searched on the basis of environmental geometry alone (i.e., at a corner with a short side of the room on the left and the long side of the room on the right).

(Montello, 1993). Rather than progressing from landmark to route to survey representations in sequence, Montello reviews evidence that all three of these types of representation develop in parallel, but at different rates. Moreover, there are large individual differences in the rate at which these representations are constructed, and some people may never construct true survey representations. Ishikawa and Montello (2006) led participants on the same routes through a novel environment once a week for 10 weeks and measured their ability to estimate straight-line directions and distances between locations. Whereas some participants had almost perfect configural knowledge of the environment after one or two learning experiences, others performed at chance and did not improve over 10 learning trials.

These striking individual differences in environmental learning can be dissociated from individual differences in smaller scale tasks that involve simple object transformations (e.g., mental rotation). In a review of 12 studies, Hegarty and Waller (2005) found that the median correlation between small-scale spatial abilities and measures of learning the layout of new environments had correlations exceeding .3 only in two studies. Correlational studies are limited in that the correlation between two variables reflects variance specific to each of the tasks, as well as error variance. Using structural equation modeling to control for these sources of variance, Hegarty et al. (2006) demonstrated that object-based spatial abilities and the ability to learn the layout of an environment are partially dissociated. Object-based spatial ability was significantly more related to learning from media (video and virtual environments) than to learning from direct experience. Moreover, learning from direct experience in an environment was somewhat distinct from learning from media (see also Liben, Myers, & Christensen, 2010). It was concluded that object-based and environmental spatial abilities share the ability to encode spatial information from visual input, maintain this information in memory, and make inferences from this information, but learning from direct experience also involves the sensing and integration of self-motion cues.

STRATEGIES IN NAVIGATION AND WAYFINDING

While researchers have isolated different navigational cues in controlled experiments, in most everyday navigation and wayfinding situations multiple cues are available, including cues about the environmental geometry, salient features, and self-motion

cues. An interesting question raised in recent spatial cognition research is the degree to which people weigh and combine these different cues in navigation and wayfinding tasks. While these questions have been asked most prominently in the debate over the use of geometry versus landmarks for reorientation, discussed earlier, researchers have begun to study how people integrate these cues in other tasks such as spatial updating (Kelly, McNamara, Bodenheimer, Carr, & Rieser, 2009) and are beginning to identify the neural structures that encode and combine environmental geometry and landmark information (Doeller, King, & Burgess, 2008). Other researchers have examined how the integration of landmarks and self-motion cues are integrated. For example, both children and adults can use both landmarks and self-motion cues to place an object in its original position in an environment, but when these cues are put in conflict, only adults integrate the two types of cues (Nardini, Jones, Bedford, & Braddick, 2008). In another study, adults depended on landmarks to walk a shortcut in a virtual environment, but they fell back on coarse survey knowledge, constructed from self-motion perception, when landmarks were unreliable (Foo, Warren, Duchon, & Tarr, 2005).

Related research has revealed that different individuals can have preferences for route-based versus survey-based navigation strategies, and intriguingly these include qualitative differences in the environmental cues and strategies used by women and men. Women more often report navigating on the basis of local landmarks and familiar routes, whereas men report using cardinal directions, environmental geometry, and metric distances (Grön, Wunderlich, Spitzer, Tomczak, & Riepe, 2000; Lawton, 1994). While women do not differ from men in dependence on, or ability to use landmarks, they depend less on geometry when reorienting to an environment (Sandstrom, Kaufman, & Huettel, 1998) and also require more environmental cues to remain oriented to an environment (Kelly et al., 2009). Strategy choice can also depend on factors such as the demands of different navigation tasks and the information available to the navigator (Hölscher, Büchner, Meilinger, & Strube, 2009). An interesting question for future research is whether the best navigators are those who switch flexibly between different strategies, depending on what is optimal in a given situation.

In summary, orientation and navigation in large-scale space depend on both visual information about

the shape of the environment and the features of the objects therein, and information from other senses, notably those involved in perception of self-motion. Research in spatial cognition at the environmental scale has revealed the types of representations that people construct from the various sensory cues, how these representations are learned, and how they are processed to guide orientation, navigation, and wayfinding. Questions of current interest include discerning the relative roles of environmental geometry, salient visual features in the environment, and self-motion cues in different aspects of orientation and wayfinding, how these sources of information are weighted and combined in different situations and in using different navigation strategies, and the brain bases of orientation and navigation.

Using Space to Think

In addition to thinking about space, at the scale of objects and environments, visuospatial thinking includes situations in which we use spatial representations to think about other entities, both abstract and concrete. Using space to think, in this way, includes metaphorical uses of space in language, thought, and graphics.

Spatial Metaphors

In general, a metaphor occurs when a source domain, which is concrete or familiar, is used to reason about a target domain, which is abstract or unfamiliar. In their seminal work on this topic, Lakoff and Johnson (1980) proposed that metaphors result directly from the nature of the body and our experience grounded in everyday events, and that metaphors are fundamentally spatial (Johnson & Lakoff, 2002; Lakoff & Nunez, 2000; Wilson & Gibbs, 2007). For example, consider the metaphor that relates up with good and down with bad. It is common to associate vertical space in this way. Phrases such as “on the top of his game” express positive and desirable conditions, while phrases such as “down in the dumps” or “downtrodden” expresses negative or undesirable conditions. Notably, this division of space is consistent across cultures and languages, which has been proposed as evidence that it is physically rooted in our experience of standing erect with pride or slumping when depressed.

The relationship between abstract concepts and space also contributes to our ability to reason about the magnitude of numbers. When asked to compare two numbers, people are faster when responding with their left hand for lower numbers and

their right hand for higher numbers (Dehaene, Bossini, & Giraux, 1993). This is known as the SNARC effect (for Spatial-Numerical Association of Response Codes). It is independent of handedness, but it is influenced by direction of writing. For example, it is reversed for Iranian participants, who write from right to left. The SNARC effect suggests that we naturally map the abstract concept of number onto space, with the direction of mapping being influenced by culture.

In a similar manner, Boroditsky and colleagues (Boroditsky, 2000, 2001; Boroditsky & Ramscar, 2002) have explored the metaphorical relationship between time and space. Time is typically mapped onto the horizontal dimension of space. This metaphorical relationship between space and time can also be expressed differently, depending on whether the observer imagines himself or herself as moving through time, or whether time is imagined as moving toward the observer (Boroditsky & Ramscar, 2002; Gentner, Imai, & Boroditsky, 2002). For example, if you had a Wednesday meeting, which was moved forward 2 days, then the new meeting day might be Friday if you imagined yourself moving through time but Monday if you imagined time moving toward you. Adoption of one or the other structural metaphor for time is affected by valence, with negative events resulting in adoption of the metaphor of time moving toward a stationary observer and positive events resulting in adoption of the metaphor of the observer moving through time (Margolies & Crawford, 2008). Finally, valence can also affect right/left mappings of time to horizontal space. Right-hand dominant participants judge an item on the right to be more preferable than an item on the left and this pattern reverses for left-hand dominant participants (Casasanto, 2008, 2009). In essence, affect may have an influence on our metaphorical mapping of time to space, which may extend to how we subjectively perceive and reason about time.

Visual and Spatial Representations in Deductive Reasoning

There has been much controversy about the functionality of visual and spatial representations in deductive reasoning (see Evans, Chapter 8). Some researchers have claimed that reasoning problems are solved by constructing spatial mental models of the contents of the problems (e.g., Huttenlocher, 1968; Johnson-Laird, 1983; see Johnson-Laird, Chapter 9), whereas others have claimed that the

fundamental representations underlying reasoning are propositional (Clark, 1969; Rips, 1994). As with most controversies of this type, neither of these extreme characterizations is likely to account for all forms of reasoning (cf. Goel, 2007). However, there is now considerable evidence from verbal protocol studies (e.g., Egan & Grimes-Farrow, 1982), dual-task studies (Vandierendonck & De Vooght, 1997), individual-difference studies (e.g., Sternberg, 1980), and brain imaging studies (e.g., Goel & Dolan, 2001; Knauff, Fangmeier, Ruff, & Johnson-Laird, 2003) that spatial representations, in addition to verbal representations, are functional in deductive reasoning. It is therefore more informative to consider the circumstances under which spatial representations are used and the nature of the spatial representations that are functional in reasoning.

FORM OF REASONING

The strongest evidence for spatial representations in reasoning is during the solution of linear syllogisms (e.g., “John is smarter than Bob, Tom is dumber than Bob, John is smarter than Tom”). De Soto, London, and Handel (1965) first proposed that when solving these reasoning problems, people form spatial representations in which they imagine the objects in the problem (e.g., John, Bob, and Tom) in a linear array in order of the relation (in this case, smartness). Huttenlocher (1968) developed this theory and likened the mental processes of constructing and manipulating mental spatial representations to constructing and moving real objects. Although purely linguistic models could also account for the solution of linear syllogisms (Clark, 1969), further studies suggested that spatial representations are functional, if not essential in linear reasoning (Shaver, Pierson, & Lang, 1975), and individual difference studies suggested that most people were best fit by a model that assumed both verbal and spatial representations (Sternberg, 1980).

In contrast with linear reasoning, other forms of reasoning such as categorical syllogisms and conditionals may be less naturally dependent on spatial representations, although spatial representations can and are often used for these types of reasoning. For example, Euler’s circles offer a visuospatial representation that is effective for solving categorical syllogisms (Stenning & Oberlander, 1995), teaching people diagrammatic representations can improve deductive reasoning with double disjunctions (Bauer & Johnson-Laird, 1993), and conditional reasoning about spatial content depends on

spatial working memory (Duyck, Vandierendonck, & De Vooght, 2003). Mental models (Johnson-Laird, 1983) are also often assumed to be visuospatial in nature.

CONTENT BEING REASONED ABOUT

Research on the effects of content on reasoning has provided insights into the nature of the spatial representations in reasoning. Although initial theories suggested very abstract spatial representations (Huttenlocher, 1968), the representations underlying reasoning were often assumed to be visuospatial images, leading to the prediction that relations that are easier to visualize are also easier to reason about (e.g., Shaver et al., 1975). This prediction was not consistently supported. In contrast, Knauff and Johnson-Laird (2002) found that when problems contain highly visual content (e.g., the dog is cleaner than the bird), reasoning is actually less efficient than when the content is spatial (e.g., the dog is in front of the bird) or abstract (e.g., the dog is smarter than the bird). This is known as the visual impedance effect. The problem with earlier research on the spatial imagery theory is that it confounded “easy to visualize” with “easy to represent as a spatial array.” Knauff and colleagues have argued that when the relation in the premises is highly visual, the construction of an abstract spatial representation of the relation is impeded by irrelevant visual detail that automatically comes to mind. In contrast, both spatial relations and abstract relations are readily mapped onto abstract spatial representations without such interference. In support of this view, a brain imaging study showed that only problems with visual relations activated vision-related areas in occipital cortex, whereas reasoning with all relations (visual, spatial, and abstract) activated a bilateral parietal-frontal system suggestive of spatial representations (Knauff et al., 2003).

CHARACTERISTICS OF THE INDIVIDUAL

It is also likely that individual differences between people affect the degree to which they use visuospatial representations in reasoning and the nature of those representations (cf. Kozhevnikov, Hegarty, & Mayer, 2002). While most participants in Sternberg’s (1980) study were best fit by a model of linear reasoning that assumed both verbal and spatial representations, a minority had a better fit to a model using verbal representations alone or spatial representations alone. Egan and Grimes-Farrow (1982) found that people could be classified as “abstract

directional thinkers,” who constructed mental orderings of the terms in linear reasoning problems, or less successful “concrete properties thinkers,” who represented the terms as detailed mental images that included physical properties of the objects (roughness, darkness, etc.). Knauff and May (2006) found that congenitally blind individuals do not show the visual impedance effect, and they argued that because blind people have never experienced visual detail, they do not represent visual properties in transitive inferences like this. Finally, dyslexics are often characterized as highly visual thinkers and are more subject to the visual impedance effect than other students (Bacon, Handley, & Newstead, 2003).

In summary, there is evidence for spatial representations in deductive reasoning, especially for linear syllogisms and when the content is either spatial or easily mapped onto spatial representations. However, there are individual differences in the extent to which people rely on spatial representations in their reasoning and the representations that are most functional in reasoning are very schematic, without visual detail.

Thinking With Diagrams, Graphs, and Maps

People also use space to think when they use graphical displays, such as diagrams, maps, and graphs. These are visuospatial external representations that can represent objects, events, or more abstract information. Graphical displays can be categorized based on the relation between the representation and its referent and how space is used to convey meaning. The first category consists of iconic displays. In iconic displays, space on the page represents space in the world and the properties displayed (shape, color, etc) are also visible properties of the referent (e.g., a road map, a diagram of human heart). These types of displays are ancient (cave paintings are examples) and used by all cultures (Tversky, 2005).

Second, relational displays are metaphorical in that they represent entities that do not have spatial extent or visible properties (e.g., when an organization chart shows the hierarchy of positions in a business, or a graph shows the price of oil over time). In these displays, visual and spatial properties represent entities and properties that are not necessarily visible or distributed over space. Visuospatial variables, such as color, shape, and location are the *representing* dimensions of the display (Bertin, 1983). The *represented* dimensions can be any category or quantity. These

types of displays are a relatively recent invention. For example the invention of the graph is attributed to Playfair in the 18th century (Wainer, 2005).

Some visual displays, including statistical maps or geospatial displays, are hybrids of these two types in that they represent both visual information and nonvisual (but spatially distributed) information. For example, a politician might examine the pattern of states that voted Republican or Democrat in the last U.S. presidential election or a researcher might examine a functional magnetic resonance image (fMRI) of activity across the brain during a cognitive task. In these displays, there is a direct mapping between space in the representation and space in the referent. However, in this case nonvisual properties are “visualized,” that is, represented by visual variables, such as color, shape, and shading, so they are also partly metaphorical.

HOW GRAPHICAL DISPLAYS ENHANCE THINKING

Graphical displays can enhance visuospatial thinking in several ways. They store information externally, freeing up working memory for other aspects of thinking (Card et al., 1999; Scaife & Rogers, 1996). They can organize information by indexing it spatially, reducing search and facilitating integration of related information (Larkin & Simon, 1987; Wickens & Carswell, 1995). These displays can also allow the offloading of cognitive processes onto automatic perceptual processes (Scaife & Rogers, 1996), or visual routines (Ullman, 1984), that abstract task-relevant properties from the display, for example, when a line in a graph reveals a linear relationship between variables (Shah, Friedman, & Vekiri, 2005). When a display is interactive, people can offload internal mental computations on external manipulations of the display itself (Card et al., 1999; Kirsh, 1997). However, although graphical displays can enhance thinking in all of these ways, this does not mean that their use is easy or transparent. In the remainder of this section we review representative studies of what is sometimes called representational or metarepresentational competence (diSessa, 2004), that is, ability to produce appropriate graphical displays, choose the best display for a given task, understand graphical displays, and interact appropriately with these displays.

PRODUCING GRAPHICAL DISPLAYS

The graphical displays that people spontaneously produce often reflect and reveal spatial metaphors.

For example, Tversky, Kugelmass, and Winter (1991) asked children to place stickers on a page to represent spatial, temporal, quantitative, and preference dimensions. They might be asked to place stickers representing breakfast, lunch, and dinner or representing their least favorite food, a food they like, and their favorite food. Most children placed the stickers in a line while maintaining the ordinal relations between the items (e.g., placing a sticker for breakfast on the left, lunch in the middle, and dinner on the right), indicating that they naturally mapped more abstract relations to space. They mapped temporal dimensions to space at an earlier age than they mapped quantitative and preference dimensions, and their mappings were affected by writing order in their cultures (for example, in the breakfast-lunch-dinner example, Arabic children, who write right to left, typically placed the sticker for breakfast on the right, lunch in the middle, and dinner on the left). Tversky (2011) argues that there are natural mappings between graphic forms and their meanings and between spatial arrangements of these forms and their meanings. For example, lines are spontaneously used to represent connections, circles to indicate cyclic processes, the horizontal dimension is naturally mapped to time, and the vertical dimension is more naturally mapped to importance or other evaluative dimensions.

In the case of more iconic displays, educators and developmental psychologists have identified children's ability to create appropriate representations, for example, in graphing motion and mapping terrain (Azevedo, 2000; diSessa, 2004; Sherrin, 2000). However, these researchers also point to limitations in this natural competence. For example, children show a strong preference for realistic representations, even when less realistic representations are more effective. There are also individual differences in ability to produce the most effective representation for a problem, especially when this representation is less realistic. For example, Hegarty and Kozhevnikov (1999) found that when solving mathematical problems, the most successful students abstracted the essential information in the problem and represented it as a schematic diagram, whereas less successful students tried to imagine irrelevant visual aspects of the objects described in the problems.

CHOOSING GRAPHICAL DISPLAYS

Another aspect of metarepresentational competence is the ability to choose the best format of external representation for a given task. In research

on relational spatial displays (matrices, networks, and hierarchies), Novick and colleagues (Novick, 2001; Novick & Hurley, 2001) found a high degree of competence among college students to match the structure of a problem to a type of diagram. They argued that people have schemas that include applicability conditions for the different types of diagrams, and they found that college students were often able to articulate these conditions.

In contrast, with more iconic and spatial displays, people show less metarepresentational competence. Undergraduate students show a strong preference for more complex iconic displays that resemble their referents by adding animation, three-dimensional information, detail, and realism (Hegarty, Smallman, Stull, & Canham, 2009). However, in fact, these display enhancements do not always improve task performance and can even impede it (e.g., Khooshabeh & Hegarty, 2010; Tversky, Morrison, & Betrancourt, 2002; Zacks, Levy, Tversky, & Schiano, 1998). Similarly, U.S. Navy experts prefer realistic three-dimensional rendered icons of ships over less realistic, more abstract symbols in their tactical displays, but they perform better with icons that pare down realism to maximize discriminability (Smallman, St. John, Oonk, & Cowen, 2001). Finally, both expert meteorologists and naïve students choose to work with weather maps that display more variables and look more realistic, but they perform faster and more accurately with simpler displays (Hegarty et al., 2009). In summary, people prefer displays that simulate the real world with greater fidelity (Scaife & Rogers 1996; Smallman & Cook, 2011; Smallman & St. John, 2005), but they are often better served by simpler, more abstract displays.

UNDERSTANDING GRAPHICAL DISPLAYS

Models of graphics comprehension propose the following three component processes in understanding graphical displays. First, users must encode the visual features of the display (e.g., lines of different slopes in a line graph). Next, they must map these onto the conceptual relationships that they convey (e.g., an upwardly sloping line shows an increasing quantity). Finally, they need to relate these conceptual relationships to the referents of the graphs (e.g., an upwardly sloping line represents an increase in the value of some stock) (Bertin, 1983; Carpenter & Shah, 1998; Pinker, 1990). Understanding a graphic can also include making inferences from the information in the display, based on the individual's prior domain knowledge or other inference rules that can

operate on the internal representation (Trickett & Trafton, 2006). Thus, graphics comprehension involves interaction between bottom-up perceptual processes of encoding information and top-down processes of applying graph schemas (Pinker, 1990; Ratwani & Trafton, 2008) and domain knowledge.

The way in which information is displayed graphically can have powerful effects on how it is interpreted and processed, providing evidence for bottom-up influences of display design. People interpret the same data differently, depending on whether they are presented in pie charts or bar graphs (Cleveland & McGill, 1986; Simkin & Hastie, 1986), bar graphs or line graphs (Shah, Mayer, & Hegarty, 1999), and which variables are assigned to the x and y axes (Gattis & Holyoak, 1996; Peebles & Cheng, 2003;

Shah & Carpenter, 1995). Effects like these can often be traced to the Gestalt principles of perceptual organization, which determine which elements of displays are grouped and can be compatible or incompatible with the tasks to be carried out with a display. Line graphs facilitate comparisons for the variable plotted on the x axis (see Fig. 31.6a) because the lines group data points as a function of this variable, reflecting the Gestalt principle of good continuation (Shah & Friedman, 2011). In contrast, bar graphs facilitate comparisons between the variable shown in the legend because the bars comparing data points with respect to this variable are closer, reflecting the Gestalt principle of proximity (see Fig. 31.6b). Similarly, Novick and Catley (2007) found differences in comprehension of two types of common hierarchical diagrams

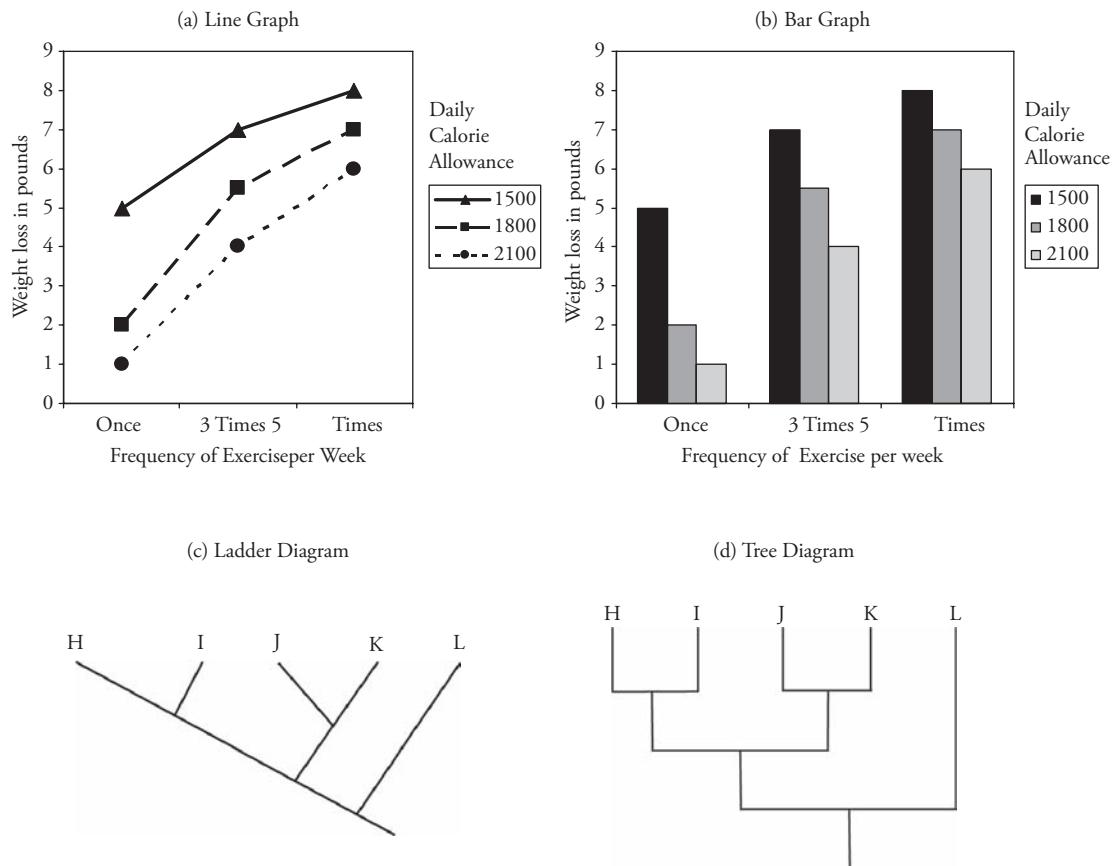


Fig. 31.6 Examples of graphics that show the same data but are processed differently. (a) and (b) are graphs of the same fictitious data showing weight loss as a function of frequency of exercise and daily calorie allowance. The line graph (a) facilitates comprehension of the effects of exercise, because the lines group data points as a function of this variable, whereas the bar graph (b) facilitates comprehension of the effects of calorie allowance because the bars comparing data points with respect to this variable are closer, reflecting the principle of proximity. (c) and (d) are examples of two types of common hierarchical diagrams (cladograms) used to display inheritance relationships in evolutionary biology. Novick and Catley (2007) found that comprehension of the ladder format (c) was more difficult than comprehension of the tree format (d), because the Gestalt principle of good continuation hindered the understanding of the hierarchical relationships in the ladder diagram.

(cladograms) in evolutionary biology, shown in Figure 31.6c and 31.6d. In the case of the diagram in Figure 31.6c, the Gestalt principle of good continuation hindered the understanding of the hierarchical relationships.

Comprehension of graphical displays is also influenced by knowledge. Experts and novices attend to different aspects of visual displays and extract different information from these displays. Experts in meteorology attend to thematically relevant aspects of displays regardless of their visual salience, whereas novices' attention is captured by the most salient features (Lowe, 1993, 1994, 1996). Experts in chess have a perceptual encoding advantage when looking at chess diagrams, such that their visual spans are larger (Reingold, Charness, Pomplun, & Stampe, 2001). Top-down effects on graphics comprehension can be separated into effects of knowledge of the graphic conventions (or graph schema) and knowledge of the domain (Shah et al., 2005; Shah & Friedman, 2011).

Finally, bottom-up effects of display design and top-down effects of knowledge can interact. Recent studies examined this interaction in a task involving making inferences from weather maps (Canham & Hegarty, 2010; Hegarty, Canham, & Fabrikant, 2010). Bottom-up effects of display design were investigated by manipulating the number of displayed variables on the maps, or the visual salience of task-relevant versus task-irrelevant information. Top-down effects of domain knowledge were investigated by examining performance and eye fixations before and after participants learned relevant meteorological principles. Map design and knowledge interacted such that salience had no effect on performance before participants learned the meteorological principles, but after learning, participants were more accurate if they viewed maps that made task-relevant information more visually salient.

INTERACTING WITH GRAPHICAL DISPLAYS

Graphical displays increasingly give the user the opportunity to choose the display he or she will use to perform a task, to add and subtract variables to graphs or maps, rotate, pan, and zoom displays. It is often assumed that this interactivity will enhance performance, by enabling people to offload internal mental computations on external manipulations of the graphic (Card et al., 1999). However, the decision to interact and choice of how to interact with a graphic also depend on knowledge of the affordances of that type of display, the problem solver's

metacognition about which transformations of displays are useful for the given problem, and the ability to make the appropriate transformation. This knowledge cannot always be assumed. In recent studies, users were given the opportunity to rotate visual displays, to perform a cross-section task, and rotating to a particular view improved performance, but many participants did not rotate the display to this optimal view (Keehner, Hegarty, Cohen, Khooshabeh, & Montello, 2008) and performed poorly. Moreover, interactivity can have other costs. For example, Yeh and Wickens (2001) found that the time and attention costs needed to decide to subtract irrelevant variables from a display outweighed the performance decrement due to visual clutter.

In summary, studies of metaphor in language and thought reveal that people naturally map abstract concepts such as time, number, and evaluative dimensions to space. Graphical displays capitalize on these natural mappings. However, this does not mean that the use of graphics is always easy. The use of spatial displays to represent both spatial and nonspatial entities is highly conventional, and the cognitive processes necessary to construct, comprehend, and use these displays are often complex and error prone. Current issues in research on this topic include the discernment of what aspects of representational competence are more and less intuitive, and how to best design graphical displays for optimal comprehension and use.

Future Directions

Our review of visuospatial cognition in this chapter has indicated that research on this topic has been dominated by many controversies and dichotomies. Is object recognition based on matching of two-dimensional visual templates or more schematic structural descriptions? Is the experience of imagery based on analog depictive representations or more abstract propositional representations? Is reorientation to an environment based on its visual features or its overall geometry? Each of these controversies reflects a distinction between more detailed, visual, often two-dimensional representations and more schematic, spatial, often three-dimensional representations. An important future direction, already evident in the research we have reviewed, will be a shift away from either-or thinking and toward understanding how different visual and spatial representations operate in different spatial thinking tasks, and how different representations and information sources are weighted and combined.

A related topic that will be important in future research is the role of strategies in visuospatial thinking. Spatial thinking tasks can often be accomplished in different ways. For example, you can find your way back to your car at the end of the day by following a well-learned route or by taking a shortcut based on a survey representation. People can solve chemistry and mechanics problems by mental simulation or by applying rules. More research is needed to examine the determinants of strategy choice in spatial thinking, which will probably include aspects of the situation and aspects of the individual (their expertise, native talents, etc.). A central question that this research will continue to address is the extent to which different methods of solving spatial problems are truly “strategic,” that is, under cognitive control, or determined by more hard-wired modular cognitive systems specialized for solving different spatial problems.

Although our review indicates that visuospatial thinking is pervasive and important in both everyday life and in many scientific disciplines, it is notable that visuospatial thinking is rarely explicitly taught. A recent National Research Council report (2006) claimed that spatial intelligence is “not just under-supported but underappreciated, undervalued, and therefore underinstructed” (p. 5) and called for a commitment to the development of spatial thinking across all areas of the school curriculum. There are important questions about the degree to which spatial thinking skills can be taught that are just beginning to be studied. For example, there is now evidence that mental rotation and aspects of spatial attention can be improved with practice (e.g., Kail, 1986; Wright, Thompson, Ganis, Newcombe, & Kosslyn, 2008) and even by experience playing video games (e.g., Feng, Spence, & Pratt, 2007; Green, Li, & Bavelier, 2010; Terlecki, Newcombe, & Little, 2008), that the effects of training can transfer to other spatial tasks (Green et al., 2010; Wright et al., 2008), and that the effects of training endure (Terlecki et al., 2008). But in training spatial thinking, what exactly should we instruct?

If we are to be most effective in fostering spatial thinking, we need to identify the basic components of this form of thinking so that training can be aimed at these fundamental components. To date, research on visuospatial thinking has been dominated by in-depth studies of a small number of paradigms such as mental rotation and reorientation in rectangular rooms. A final (or perhaps preliminary) goal we identify for future research on visuospatial cognition (or spatial

cognition more generally) is the development of a taxonomy of spatial thinking tasks that is grounded in an understanding of the different processes that make up visuospatial reasoning (see Wiener, Büchner, & Hölscher, 2009 for a preliminary taxonomy of wayfinding tasks). This will involve looking beyond the laboratory into the world to better characterize the spatial thinking challenges that people face in everyday life and in their professions, but also studying these under controlled conditions using virtual environment and brain imaging technologies as well as more traditional methods. With the new recognition of the importance of spatial thinking, this promises to be an important and exciting endeavor.

Acknowledgments

The preparation of this chapter was funded in part by grant DRL-1008650 from the National Science Foundation. We thank Rob Goldstone and Keith Holyoak for comments on an earlier version of the chapter.

References

- Avraamides, M., Loomis, J. M., Klatzky, R. L., & Golledge, R. G. (2004). Functional equivalence of spatial representations derived from vision and language: Evidence from allocentric judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 801–814.
- Azevedo, F. S. (2000). Designing representations of terrain: A study in meta-representational competence. *The Journal of Mathematical Behavior*, 19, 443–480.
- Bacon, A. M., Handley, S. J., & Newstead, S. E. (2003). Individual differences in strategies for syllogistic reasoning. *Thinking and Reasoning*, 9, 133–168.
- Baddeley, A. D., & Lieberman, K. (1980). Spatial working memory. In R. S. Nickerson (Ed.), *Attention and performance* (pp. 521–539). Hillsdale, NJ: Erlbaum.
- Bauer, M. I., & Johnson-Laird, P. N. (1993). How diagrams can improve reasoning. *Psychological Science*, 4(6), 372–378.
- Berti, A., & Frassinetti, F. (2000). When far becomes near: Remapping of space by tool use. *Journal of Cognitive Neuroscience*, 12, 415–420.
- Bertin, J. (1983). *Semiology of graphics: Diagrams, networks, maps* (W. Berg, Trans.). Madison: University of Wisconsin Press.
- Bethell-Fox, C. E., & Shepard, R. N. (1988). Mental rotation: Effects of stimulus complexity and familiarity. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 12–23.
- Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94, 115–147.
- Bisiach, E., & Luzzatti, C. (1978). Unilateral neglect of representational space. *Cortex*, 14, 129–133.
- Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition*, 75, 1–28.
- Boroditsky, L. (2001). Does language shape thought? Mandarin and English speakers’ conceptions of time. *Cognitive Psychology*, 43, 1–22.
- Boroditsky, L., & Ramscar, M. (2002). The roles of body and mind in abstract thought. *Psychological Science*, 13, 185–189.

- Byrne, P., Becker, S., & Burgess, N. (2007). Remembering the past and imagining the future: A neural model of spatial memory and imagery. *Psychological Review*, 114(2), 340–375.
- Canham, M., & Hegarty, M. (2010). Effects of knowledge and display design on comprehension of complex graphics. *Learning and Instruction*, 20, 155–166.
- Card, S. K., Mackinlay, J. D., & Schneiderman, B. (1999). *Readings in information visualization: Using vision to think*. San Francisco, CA: Morgan Kaufmann Publishers.
- Carlson, T. A., Alvarez, G., Wu, D., & Verstraten, F. A. J. (2010). Rapid assimilation of external objects into the body schema. *Psychological Science*, 21, 1000–1005.
- Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied*, 4, 75–100.
- Casasanto, D. (2008). Who's afraid of the big bad Whorf? Crosslinguistic differences in temporal language and thought. *Language Learning*, 58, 63–79.
- Casasanto, D. (2009). Embodiment of abstract concepts: Good and bad in right- and left-handers. *Journal of Experimental Psychology: General*, 138, 351–367.
- Chambers, D. (1997). Are images vivid pictures in the mind? *PsychCRITIQUES*, 42, 613–614.
- Chambers, D., & Reisberg, D. (1985). Can mental images be ambiguous? *Journal of Experimental Psychology: Human Perception and Performance*, 11, 317–328.
- Chambers, D., & Reisberg, D. (1992). What an image depicts depends on what an image means. *Cognitive Psychology*, 24, 145–174.
- Cheng, K. (1986). A purely geometric module in the rats spatial representation. *Cognition*, 23, 149–178.
- Cheng, K., & Newcombe, N. (2005). Is there a geometric module for spatial reorientation? Squaring theory and evidence. *Psychological Bulletin and Review*, 12, 1–23.
- Cheng, K., Shettleworth, S. J., Huttenlocher, J., & Rieser, J. J. (2007). Bayesian integration of spatial information. *Psychological Bulletin*, 133, 625–637.
- Clark, H. H. (1969). Linguistic processes in deductive reasoning. *Psychological Review*, 76, 387–404.
- Cleveland, W. S., & McGill, R. (1986). An experiment in graphical perception. *International Journal of Man-Machine Studies*, 25, 491–500.
- Conson, M., Sarà, M., Pistoia, F., & Trojano, L. (2009). Action observation improves motor imagery: Specific interactions between simulative processes. *Experimental Brain Research*, 199, 71–81.
- Cooper, L. A. (1976). Demonstration of a mental analog of an external rotation. *Perception and Psychophysics*, 19, 296–302.
- Cooper, L. A., & Podgorny, R. (1976). Mental transformations and visual comparison processes: Effects of complexity and similarity. *Journal of Experimental Psychology: Human Perception and Performance*, 2, 503–514.
- Cooper, L. A., & Shepard, R. N. (1984). Turning something over in one's mind. *Scientific American*, 251, 106–114.
- Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General*, 122, 371–396.
- De Soto, C. B., London, M., & Handel, S. (1965). Social reasoning and spatial paralogic. *Journal of Personality and Social Psychology*, 2, 513–520.
- diSessa, A. A. (2004). Metarepresentation: Native competence and targets for instruction. *Cognition and Instruction*, 22, 293–331.
- Diwadkar, V. A., & McNamara, T. P. (1997). Viewpoint dependence in scene recognition. *Psychological Science*, 8(4), 302–307.
- Doeller, C. F., King, J. A., & Burgess, N. (2008). Parallel striatal and hippocampal systems for landmarks and boundaries in spatial memory. *Proceedings of the National Academy of Sciences USA*, 105, 5915–5920.
- Duyck, W., Vandierendonck, A., & De Vooght, G. (2003). Conditional reasoning with spatial content requires visuo-spatial working memory. *Thinking and Reasoning*, 9, 267–287.
- Easton, R. D., & Sholl, M. J. (1995). Object-array structure, frames of reference, and retrieval of spatial knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 483–500.
- Egan, D. E., & Grimes-Farrow, D. D. (1982). Differences in mental representations spontaneously adopted for reasoning. *Memory and Cognition*, 10, 297–307.
- Emery, N. J., & Clayton, N. S. (2004). The mentality of crows: Convergent evolution of intelligence in corvids and apes. *Science*, 306, 1903–1907.
- Epstein, R. (2008). Parahippocampal and retrosplenial contributions to human spatial navigation. *Trends in Cognitive Sciences*, 12, 388–396.
- Farah, M. J. (1988). Is visual imagery really visual? Overlooked evidence from neuropsychology. *Psychological Review*, 95, 307–317.
- Farrell, M. J., & Robertson, I. (1998). Mental rotation and the automatic updating of body-centered spatial relationships. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 227–233.
- Feng, J., Spence, I., & Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. *Psychological Science*, 18, 850–855.
- Finke, R. A. (1989). *Principles of mental imagery*. Cambridge, MA: MIT Press.
- Finke, R. A., & Pinker, S. (1983). Directional scanning of remembered visual patterns. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9, 398–410.
- Finke, R. A., Pinker, S., & Farah, M. J. (1989). Reinterpreting visual patterns in mental imagery. *Cognitive Science: A Multidisciplinary Journal*, 13, 51–78.
- Folk, M. D., & Luce, R. D. (1987). Effects of stimulus complexity on mental rotation rate of polygons. *Journal of Experimental Psychology: Human Perception and Performance*, 13, 395–404.
- Foo, P., Warren, W. H., Duchon, A., & Tarr, M. J. (2005). Do humans integrate routes into a cognitive map? Map- versus landmark-based navigation of novel shortcuts. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 31, 195–215.
- Foster, D. H., & Gilson, S. J. (2002). Recognizing novel three-dimensional objects by summing signals from parts and views. *Proceedings of the Royal Society B: Biological Sciences*, 269, 1939–1947.
- Friedman, A., Brown, N. R., & McGaffey, A. P. (2002). A basis for bias in geographical judgments. *Psychonomic Bulletin and Review*, 9, 151–159.
- Friedman, A., Sptch, M. L., & Ferrey, A. (2005). Recognition by humans and pigeons of novel views of 3-D objects and their photographs. *Journal of Experimental Psychology: General*, 134, 149–162.
- Gallistel, C. R. (1990). *The organization of learning*. Cambridge, MA: MIT Press.

- Galton, F. (1883). *Inquiries into human faculty and its development*. Bristol, England: Thoemmes Press.
- Gardner, H. (2004). *Frames of mind: The theory of multiple intelligences*. New York: Basic Books.
- Gattis, M., & Holyoak, K. J. (1996). Mapping conceptual to spatial relations in visual reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 231–239.
- Gauthier, I., & Tarr, M. J. (1997). Orientation priming of novel shapes in the context of viewpoint-dependent recognition. *Perception*, 26, 51–73.
- Gentner, D., Imai, M., & Boroditsky, L. (2002). As time goes by: Evidence for two systems in processing space-time metaphors. *Language and Cognitive Processes*, 17, 537–565.
- Goel, V. (2007). Anatomy of deductive reasoning. *Trends in Cognitive Sciences*, 11, 435–441.
- Goel, V., & Dolan, R. J. (2001). Functional neuroanatomy of three-term relational reasoning. *Neuropsychologia*, 39, 901–909.
- Goldstone, R. L., Landy, D., & Son, J. Y. (2010). The education of perception. *Topics in Cognitive Science*, 2, 265–284.
- Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neuroscience*, 15, 20–25.
- Graf, M. (2006). Coordinate transformations in object recognition. *Psychological Bulletin*, 132, 920–945.
- Graf, M. (2010). Categorization and object shape. In B. M. Glatzeder, V. Goel, & A. von Müller (Eds.), *Towards a theory of thinking: Building blocks for a conceptual framework* (pp. 73–102). Berlin, Germany: Springer-Verlag.
- Graf, M., Kaping, D., & Bühlhoff, H. H. (2005). Orientation congruency effects for familiar objects: Coordinate transformations in object recognition. *Psychological Science*, 16, 214–221.
- Green, S., Li, R., & Bavelier (2010). Perceptual learning during action video game playing. *Topics in Cognitive Science*, 2, 202–216.
- Grön, G., Wunderlich, A. P., Spitzer, M., Tomczak, R., & Riepe M. W. (2000). Brain activation during human navigation: Gender-different neural networks as substrate of performance. *Nature Neuroscience*, 3, 404–408.
- Halligan, P. W., Fink, G. R., Marshall, J. C., & Vallar, G. (2003). Spatial cognition: Evidence from visual neglect. *Trends in Cognitive Sciences*, 7, 125–133.
- Hayward, W. G. (2003). After the viewpoint debate: Where next in object recognition? *Trends in Cognitive Science*, 7, 425–427.
- Hegarty, M. (1992). Mental animation: Inferring motion from static displays of mechanical systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 1084–1102.
- Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Science*, 18, 280–285.
- Hegarty, M. (2010). Components of spatial intelligence. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 265–297). San Diego, CA: Elsevier Academic Press.
- Hegarty, M., Canham, M., & Fabrikant, S. I. (2010). Thinking about the weather: How display salience and knowledge affect performance in a graphic inference task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36, 37–53.
- Hegarty, M., & Kozhevnikov, M. (1999). Types of visuospatial representations and mathematical problem solving. *Journal of Educational Psychology*, 91, 684–689.
- Hegarty, M., Mayer, S., Kriz, S., & Keehner, M. (2005). The role of gestures in mental animation. *Spatial Cognition and Computation*, 5, 333–356.
- Hegarty, M., Montello, D. R., Richardson, A. E., Ishikawa, T., & Lovelace, K. (2006). Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. *Intelligence*, 34, 151–176.
- Hegarty, M., Smallman, H., Stull, A. T., & Canham, M. (2009). Naïve cartography: How intuitions about display configuration can hurt performance. *Cartographica*, 44, 171–186.
- Hegarty, M., & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, 32, 175–191.
- Hegarty, M., & Waller, D. (2005). Individual differences in spatial ability. In P. Shah (Ed.), *The Cambridge handbook of visuospatial thinking* (pp. 121–169). New York: Cambridge University Press.
- Hermer, L., & Spelke, E. (1994). A geometric process for spatial representation in young children. *Nature*, 370, 57–59.
- Hermer-Vasquez, L., Spelke, E., & Katsnelson, A. (1999). Sources of flexibility in human cognition: Dual task studies of space and language. *Cognitive Psychology*, 39, 3–36.
- Hirtle, S. C., & Jonides, J. (1985). Evidence of hierarchies in cognitive maps. *Memory and Cognition*, 13, 208–217.
- Hintzman, D. L., O'Dell, C. S., & Arndt, D. R. (1981). Orientation in cognitive maps. *Cognitive Psychology*, 13, 149–206.
- Hölscher, C., Büchner, S. J., Meilinger, T., & Strube, G. (2009). Adaptivity of wayfinding strategies in a multi-building ensemble: The effects of spatial structure, task requirements and metric information. *Journal of Environmental Psychology*, 29, 208–219.
- Holyoak, K. J., & Mah, W. A. (1982). Cognitive reference points in judgments of symbolic magnitude. *Cognitive Psychology*, 14, 328–352.
- Hummel, J. E. (2003). The complementary properties of holistic and analytic representations of object shape. In M. Peterson & G. Rhodes (Eds.), *Perception of faces, objects, and scenes: Analytic and holistic processes* (pp. 212–234). New York: Oxford University Press.
- Humphreys, G. W., & Riddoch, M. J. (1984). Routes to object constancy: Implications from neurological impairments of object constancy. *Quarterly Journal of Experimental Psychology A*, 37, 493–495.
- Huttenlocher, J. (1968). Constructing spatial images: A strategy in reasoning. *Psychological Review*, 75, 550–560.
- Ishikawa, T., & Montello, D. R. (2006). Spatial knowledge acquisition from direct experience in the environment: Individual differences in the development of metric knowledge and the integration of separately learned places. *Cognitive Psychology*, 52, 93–129.
- Johnson, M., & Lakoff, G. (2002). Why cognitive linguistics requires embodied realism. *Cognitive Linguistics*, 13, 245–263.
- Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, England: Cambridge University Press.
- Jolicoeur, P. (1990). Orientation congruency effects on the identification of disoriented shapes. *Journal of Experimental Psychology: Human Perception and Performance*, 16, 351–364.
- Jolicoeur, P. (1992). Orientation congruency effects in visual search. *Canadian Journal of Psychology*, 46, 280–305.
- Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, 8, 441–480.

- Just, M. A., & Carpenter, P. A. (1985). Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability. *Psychological Review*, 92, 137–172.
- Kail, R. (1986). The impact of extended practice on rate of mental rotation. *Journal of Experimental Child Psychology*, 42, 378–391.
- Keehner, M., Hegarty, M., Cohen, C. A., Khooshabeh, P., & Montello, D. R. (2008). Spatial reasoning with external visualizations: What matters is what you see, not whether you interact. *Cognitive Science*, 32, 1099–1132.
- Kelly, J., McNamara, T. P., Bodenheimer, B., Carr, T. H., & Rieser, J. J. (2009). Individual differences in using geometric and featural cues to maintain spatial orientation: Cue quantity and cue ambiguity are more important than cue type. *Psychonomic Bulletin and Review*, 16, 176–181.
- Khooshabeh, P., & Hegarty, M. (2010). Inferring cross-sections: When internal visualizations are more important than properties of external visualizations. *Human-Computer Interaction*, 25, 119–147.
- Kirsh, D. (1997). Interactivity and multimedia interfaces. *Instructional Science*, 25, 79–96.
- Klatzky, R. L., Loomis, J. M., Beall, A. C., Chance, S. S., & Golledge, R. G. (1998). Spatial updating of self-position and orientation during real, imagined, and virtual locomotion. *Psychological Science*, 9, 293–298.
- Knauff, M., Fangmeier, T., Ruff, C. C., & Johnson-Laird, P. N. (2003). Reasoning, models and images: Behavioral measures of cortical activity. *Journal of Cognitive Neuroscience*, 4, 559–573.
- Knauff, M., & Johnson-Laird, P. N. (2002). Visual imagery can impede reasoning. *Memory and Cognition*, 10, 363–371.
- Knauff, M., & May, E. (2006). Mental imagery, reasoning, and blindness. *The Quarterly Journal of Experimental Psychology*, 59, 161–177.
- Kosslyn, S. M. (1980). *Image and Mind*. Cambridge, MA: Harvard University Press.
- Kosslyn, S. M., Alpert, N. M., Thompson, W. L., & Maljkovic, V. (1993). Visual mental imagery activates topographically organized visual cortex: PET investigations. *Journal of Cognitive Neuroscience*, 5, 263–287.
- Kosslyn, S. M., Ball, T. M., & Reiser, B. J. (1978). Visual images preserve metric spatial information: Evidence from studies of image scanning. *Journal of Experimental Psychology: Human Perception and Performance*, 4, 47–60.
- Kosslyn, S. M., Brunn, J., Cave, K. R., & Wallach, R. W. (1984). Individual differences in mental imagery ability: A computational analysis. *Cognition*, 18, 195–243.
- Kosslyn, S. M., & Thompson, W. L. (2003). When is early visual cortex activated during visual mental imagery? *Psychological Bulletin*, 129, 723–746.
- Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2006). *The case for mental imagery*. New York: Oxford University Press.
- Kosslyn, S. M., Thompson, W. L., Wraga, M., & Alpert, N. M. (2001). Imaging rotation by endogenous versus exogenous forces: Distinct neural mechanisms. *NeuroReport: For Rapid Communication of Neuroscience Research*, 12, 2519–2525.
- Kozhevnikov, M., Blazhenkova, O., & Becker, M. (2010). Trade-off in object versus spatial visualization abilities: Restriction in the development of visual-processing resources. *Psychonomic Bulletin and Review*, 17, 29–35.
- Kozhevnikov, M., & Hegarty, M. (2001). A dissociation between object-manipulation and perspective-taking spatial abilities. *Memory and Cognition*, 29, 745–756.
- Kozhevnikov, M., Hegarty, M., & Mayer, R. E. (2002). Revising the visualizer-verbalizer dimension: Evidence for two types of visualizers. *Cognition and Instruction*, 20, 47–78.
- Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago, IL: University of Chicago Press.
- Lakoff, G., & Nunez, R. (2000). *Where mathematics comes from*. New York: Basic Books.
- Larkin, J., & Simon, H. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11, 65–99.
- Lawson, R. (1999). Achieving visual object constancy across plane rotation and depth rotation. *Acta Psychologica*, 102, 221–245.
- Lawson, R., & Humphreys, G. W. (1999). The effects of view in depth on the recognition of line drawings and silhouettes of familiar objects: Normality and pathology. *Visual Cognition*, 6, 165–195.
- Lawton, C. A. (1994). Gender differences in wayfinding strategies: Relationship to spatial ability and spatial anxiety. *Sex Roles*, 30, 765–779.
- Levine, M., Marchon, I., & Hanley, G. L. (1984). The placement and misplacement of you-are-here maps. *Environment and Behavior*, 16, 139–157.
- Liben, L. S., Myers, L. J., & Christensen, A. E. (2010). Identifying locations and directions on field and representational mapping tasks: Predictors of success. *Spatial Cognition and Computation*, 10, 105–134.
- Logie, R. H. (1995). Visuo-spatial working memory. Hillsdale, NJ: Erlbaum.
- Lohman, D. F. (1988). Spatial abilities as traits, processes, and knowledge. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (pp. 181–248). Hillsdale, NJ: Erlbaum.
- Lowe, R. K. (1993). Constructing a mental representation from an abstract technical diagram. *Learning and Instruction*, 3, 157–179.
- Lowe, R. K. (1994). Selectivity in diagrams: Reading beyond the lines. *Educational Psychology*, 14, 467–491.
- Lowe, R. K. (1996). Background knowledge and the construction of a situational representation from a diagram. *European Journal of Psychology of Education*, 11, 377–397.
- Macaluso, E., & Maravita, A. (2010). The representation of space near the body through touch and vision. *Neuropsychologia*, 48, 782–795.
- Makin, T. R., Wilf, M., Schwartz, I., & Zohary, E. (2009). Amputees “neglect” the space near their missing hand. *Psychological Science*, 21, 55–57.
- Maravita, A., & Iriki, A. (2004). Tools for the body (schema). *Trends in Cognitive Sciences*, 8, 79–86.
- Margolies, S. O., & Crawford, L. E. (2008). Event valence and spatial metaphors of time. *Cognition and Emotion*, 22, 1401–1414.
- Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society B*, 200, 269–294.
- May, M. (1996). Cognitive and embodied modes of spatial imagery. *Psychologische Beiträge*, 38, 418–434.
- Montello, D. R. (1993). Scale and multiple psychologies of space. In A. U. Frank & I. Campari (Eds.), *Spatial information theory: A theoretical basis for GIS* (pp. 312–321). Berlin, Germany: Springer-Verlag.
- Mou, W., & McNamara, T. P. (2002). Intrinsic frames of reference in spatial memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 162–170.

- Muller, M., & Wehner, R. (1988). Path integration in desert ants, *Cataglyphis fortis*. *Proceedings of the National Academy of Sciences USA*, *85*, 5287–5290.
- Nardini, M., Jones, P., Bedford, R., & Braddick, O. (2008). Development of cue integration in human navigation. *Current Biology*, *18*, 689–693.
- National Research Council. (2006). *Learning to think spatially: GIS as a support system in the K-12 curriculum*. Washington, DC: National Research Council Press.
- Novick, L. R. (2001). Spatial diagrams: Key instruments in the toolbox for thought. In D. L. Medin (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 279–325). San Diego, CA: Academic Press.
- Novick, L. R., & Catley, K. M. (2007). Understanding phylogenies in biology: The influence of a Gestalt perceptual principle. *Journal of Experimental Psychology: Applied*, *13*, 197–223.
- Novick, L. R., & Hurley, S. M. (2001). To matrix, network, or hierarchy: That is the question. *Cognitive Psychology*, *42*, 158–216.
- Pani, J. R. (1993). Limits on the comprehension of rotational motion: Mental imagery of rotations with oblique components. *Perception*, *22*, 785–808.
- Pani, J. R., William, C. T., & Shippey, G. T. (1995). Determinants of the perception of rotational motion: Orientation of the motion to the object and to the environment. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 1441–1456.
- Panis, S., Vangeneugden, J., & Wagemans, J. (2008). Similarity, typicality, and category-level matching of morphed outlines of everyday objects. *Perception*, *37*, 1822–1849.
- Parsons, L. M. (1987a). Imagined spatial transformations of one's hands and feet. *Cognitive Psychology*, *19*, 178–241.
- Parsons, L. M. (1987b). Imagined spatial transformation of one's body. *Journal of Experimental Psychology: General*, *116*, 172–191.
- Parsons, L. M. (1994). Temporal and kinematic properties of motor behavior reflected in mentally simulated actions. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 709–730.
- Peebles, D., & Cheng, P. C-H. (2003). Modeling the effect of task and graphical representation on response latency in a graph-reading task. *Human Factors*, *45*, 28–46.
- Pinker, S. (1990). A theory of graph comprehension. In R. Freedle (Ed.), *Artificial intelligence and the future of testing* (pp. 73–126). Hillsdale, NJ: Erlbaum.
- Presson, C. C. (1987). The development of spatial cognition: Secondary uses of spatial information. In N. Eisenberg (Ed.) *Contemporary topics in developmental psychology* (pp. 77–112). New York: Wiley.
- Presson, C. C., & Montello, D. R. (1994). Updating after rotational and translational body movements: Coordinate structure of perspective space. *Perception*, *23*, 1447–1455.
- Previc, F. H. (1998). The neuropsychology of 3-D space. *Psychological Bulletin*, *124*, 123–164.
- Pyllyshyn, Z. W. (2003). *Seeing and visualizing: It's not what you think*. Cambridge, MA: MIT Press.
- Ratwani, R. M., & Trafton, J. G. (2008). Shedding light on the graph schema: Perceptual features versus invariant structure. *Psychonomic Bulletin and Review*, *15*, 757–762.
- Reed, C. L., Betz, R., Garza, J. P., & Roberts, R. J., Jr. (2009). Grab it! Biased attention in functional hand and tool space. *Attention, Perception, and Psychophysics*, *72*, 236–245.
- Reed, C. L., Grubb, J. D., & Steele, C. (2006). Hands up: Attentional prioritization of space near the hand. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 166–177.
- Reed, S. K. (1974). Structural descriptions and the limitations of mental images. *Memory and Cognition*, *2*, 319–336.
- Reingold, E. M., Charness, N., Pomplun, M., & Stampe, D. M. (2001). Visual span in expert chess players: Evidence from eye movements. *Psychological Science*, *12*, 48–55.
- Richardson, A. E., Montello, D., & Hegarty, M. (1999). Spatial knowledge acquisition from maps, and from navigation in real and virtual environments. *Memory and Cognition*, *27*, 741–750.
- Rieser, J. J. (1989). Access to knowledge of spatial structure at novel points of observation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 1157–1165.
- Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. Cambridge, MA: MIT Press.
- Sack, A. T., Lindner, M., & Linden, D. E. J. (2007). Object- and direction-specific interference between manual and mental rotation. *Perception and Psychophysics*, *69*, 1435–1449.
- Sadalla, E. K., Burroughs, W. J., & Staplin, L. J. (1980). Reference points in spatial cognition. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 516–528.
- Sandstrom, N. J., Kaufman, J., & Huettel, S. A. (1998). Males and females use different distal cues in a virtual environment navigation task. *Cognitive Brain Research*, *6*, 351–360.
- Scaife, M., & Rogers, Y. (1996). External cognition: How do graphical representations work? *International Journal of Human-Computer Studies*, *45*, 185–213.
- Schwartz, D. L., & Black, J. B. (1996). Shutting between depictive models and abstract rules: Induction and fallback. *Cognitive Science: A Multidisciplinary Journal*, *20*, 457–497.
- Schwartz, D. L., & Black, T. (1999). Inferences through imagined actions: Knowing by simulated doing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 116–136.
- Schwartz, D. L., & Holton, D. L. (2000). Tool use and the effect of action on the imagination. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1655–1665.
- Sekiyama, K. (1982). Kinesthetic aspects of mental representations in the identification of left and right hands. *Perception and Psychophysics*, *32*, 89–95.
- Shah, P., & Carpenter, P. (1995). Conceptual limitations in comprehending line graphs. *Journal of Experimental Psychology: General*, *124*, 337–370.
- Shah, P., & Friedman, E. G. (2011). Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *TopICS in Cognitive Science*, *3*(3), 560–578.
- Shah, P., Friedman, E. G., & Vekiri, I. (2005). The comprehension of quantitative information in graphical displays. In P. Shah & A. Miyake (Eds.), *The Cambridge handbook of visuospatial thinking* (pp. 426–476). New York: Cambridge University Press.
- Shah, P., Mayer, R. E., & Hegarty, M. (1999). Graphs as aids to knowledge construction. *Journal of Educational Psychology*, *91*, 690–702.
- Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing. *Journal of Experimental Psychology: General*, *125*, 4–27.
- Shaver, P., Pierson, L., & Lang, S. (1975). Converging evidence for the functional significance of imagery in problem solving. *Cognition*, *3*, 359–375.

- Shelton, A. L., & McNamara, T. P. (1997). Multiple views of spatial memory. *Psychonomic Bulletin and Review*, 4, 102–106.
- Shelton, A. L., & McNamara, T. P. (2001). Systems of spatial reference in human memory. *Cognitive Psychology*, 43, 274–310.
- Shepard, R. M., & Metzler, L. (1971). Mental rotation of three dimensional objects. *Science*, 191, 952–954.
- Sherrin, B. (2000). How students invent representations of motion. *Journal of Mathematical Behavior*, 19, 399–441.
- Siegel, A. W., & White, S. H. (1975). The development of spatial representations of large-scale environments. In H. W. Reese (Ed.), *Advances in child development and behavior* (Vol 10, pp. 9–55). New York: Academic Press.
- Simkin, D. K., & Hastie, R. (1986). An information-processing analysis of graph perception. *Journal of the American Statistical Association*, 82, 454–465.
- Simons, D. J., & Wang, R. F. (1998). Perceiving real-world viewpoint changes. *Psychological Science*, 9, 315–320.
- Sims, V. K., & Hegarty, M. (1997). Mental animation in the visuospatial sketchpad: Evidence from dual-task studies. *Memory and Cognition*, 25, 321–332.
- Smallman, H. S., & Cook, M. B. (2011). Naïve realism: Folk fallacies in the design and use of visual displays. *TopiCS in Cognitive Science*, 3(3), 579–608.
- Smallman, H. S., & St. John, M. (2005). Naïve realism: Misplaced faith in realistic displays. *Ergonomics in Design*, 13, 14–19.
- Smallman, H. S., St. John, M., Oonk, H. M., & Cowen, M. B. (2001). 'SYMBICONS': A hybrid symbology that combines the best elements of SYMBols and ICONS. In *Proceedings of the 45th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 110–114). Santa Monica, CA: HFES.
- Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, 10, 89–96.
- Stenning, K., & Oberlander, J. (1995). A cognitive theory of graphical and linguistic reasoning: Logic and implementation. *Cognitive Science*, 19, 97–140.
- Sternberg, R. J. (1980). Representation and process in linear syllogistic reasoning. *Journal of Experimental Psychology: General*, 109, 119–159.
- Stevens, A., & Coupe, P. (1978). Distortions in judged spatial relations. *Cognitive Psychology*, 10, 422–437.
- Steiff, M. (2007). Mental rotation and diagrammatic reasoning in science. *Learning and Instruction*, 17, 219–234.
- Steiff, M., Hegarty, M., & Dixon B. L. (2010). Alternative strategies for spatial reasoning with diagrams. In A. Goel, M. Jamnik, & N. H. Narayanan (Eds.), *Diagrammatic representation and inference (Proceedings of Diagrams 2010)*. Berlin, Germany: Springer-Verlag.
- Steiff, M., & Raje, S. (2010). Expert algorithmic and imagistic problem solving strategies in advanced chemistry. *Spatial Cognition and Computation*, 10, 53–81.
- Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, 21, 233–282.
- Terlecki, M. S., Newcombe, N. S., & Little, M. (2008). Durable and generalized effects of spatial experience on mental rotation: Gender differences in growth patterns. *Applied Cognitive Psychology*, 22, 996–1013.
- Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381, 520–522.
- Thurstone, L. L. (1938). *Primary mental abilities*. Chicago IL: University of Chicago Press.
- Trickett, S. B., & Trafton, J. G. (2006). Toward a comprehensive model of graph comprehension: Making the case for spatial cognition. In D. Barker-Plummer, R. Cox, & N. Swaboda (Eds.), *Diagrammatic representation and inference* (pp. 286–300). Berlin, Germany: Springer.
- Tversky, B. T. (1981). Distortions in memory for maps. *Cognitive Psychology*, 13, 407–433.
- Tversky, B. (2005). Visuospatial reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 209–240). New York, NY: Cambridge University Press.
- Tversky, B. (2011). Visualizing thought. *TopiCS in Cognitive Science*, 3(3), 499–535.
- Tversky, B., Agrawala, M., Heiser, J., Lee, P. U., Hanrahan, P., Phan, D., et al. (2007). Cognitive design principles for generating visualizations. In G. Allen (Ed.), *Applied spatial cognition: From research to cognitive technology* (pp. 53–73). Mahwah, NJ: Erlbaum.
- Tversky, B., Kugelmass, S., & Winter, A. (1991). Cross-cultural and developmental trends in graphic productions. *Cognitive Psychology*, 23, 515–577.
- Tversky, B., Morrison, J. B., & Betrancourt, M. (2002). Animation: Can it facilitate? *International Journal of Human-Computer Studies*, 57, 247–262.
- Twyman, A. D., & Newcombe, N. S. (2010). Five reasons to doubt the existence of a geometric module. *Cognitive Science*, 34, 1315–1356.
- Ullman, S. (1984). Visual routines. *Cognition*, 18, 97–159.
- Ungerleider, L. G., & Mishkin, M. (1982). Two cortical systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *Analysis of visual behavior* (pp. 549–586). Cambridge, MA: MIT Press.
- Vandierendonck, A., & De Vooght, G. (1997). Working memory constraints on linear reasoning with spatial and temporal contents. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 50, 803–820.
- Vanrie, J., Willems, B., & Wagemans, J. (2001). Multiple routes to object matching from different viewpoints: Mental rotation versus invariant features. *Perception*, 30, 1047–1056.
- Wainer, H. (2005). *Graphic discovery: A trout in the milk and other visual adventures*. Princeton, NJ: Princeton University Press.
- Waller, D., & Hodgson, E. (2006). Transient and enduring spatial representations under disorientation and self-rotation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32, 867–882.
- Waller, D., Montello, D. R., Richardson, A. E., & Hegarty, M. (2002). Orientation specificity and spatial updating of memories for layouts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 1051–1063.
- Wang, R. F., & Simons, D. J. (1999). Active and passive scene recognition across views. *Cognition*, 70, 191–210.
- Wang, R. F., & Spelke, E. S. (2000). Updating egocentric representations in human navigation. *Cognition*, 77, 215–250.
- Wexler, M., Kosslyn, S. M., & Berthoz, A. (1998). Motor processes in mental rotation. *Cognition*, 68, 77–94.
- Wickens, C. D., & Carswell, M. (1995). The proximity compatibility principle: Its psychological foundation and relevance to display design. *Human Factors*, 37, 473–494.
- Wiener, J. M., Büchner, S. J., & Hölscher, C. (2009). Taxonomy of human wayfinding tasks: A knowledge-based approach. *Spatial Cognition and Computation*, 9, 152–165.

- Wilson, N. L., & Gibbs, R. W., Jr. (2007). Real and imagined body movement primes metaphor comprehension. *Cognitive Science*, 31, 721–731.
- Wohlschläger, A. (2001). Mental object rotation and the planning of hand movements. *Perception and Psychophysics*, 63, 709–718.
- Wohlschläger, A., & Wohlschläger, A. (1998). Mental and manual rotation. *Journal of Experimental Psychology: Human Perception and Performance*, 24, 397–412.
- Wolbers, T., & Hegarty, M. (2010). What determines our navigational abilities? *Trends in Cognitive Sciences*, 14, 138–146.
- Wright, R., Thompson, W. L., Ganis, G., Newcombe, N. S., & Kosslyn, S. M. (2008). Training generalized spatial skills. *Psychonomic Bulletin and Review*, 15, 763–771.
- Xiao, C., Mou, W., & McNamara, T. P. (2009). Use of self-to-object and object-to-object spatial relations in locomotion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 1137–1147.
- Yeh, M., & Wickens, C. D. (2001). Attentional filtering in the design of electronic map displays: A comparison of color coding, intensity coding, and decluttering techniques. *Human Factors*, 43, 543–562.
- Yuille, J. C., & Steiger, J. H. (1982). Nonholistic processing in mental rotation: Some suggestive evidence. *Perception and Psychophysics*, 31, 201–209.
- Zacks, J. M. (2008). Neuroimaging studies of mental rotation: A meta-analysis and review. *Journal of Cognitive Neuroscience*, 20, 1–19.
- Zacks, J. M., Levy, E., Tversky, B., & Schiano, D. J. (1998). Reading bar graphs: Effects of extraneous depth cues and graphical context. *Journal of Experimental Psychology: Applied*, 4, 119–138.

Gesture in Thought

Susan Goldin-Meadow and Susan Wagner Cook

Abstract

The spontaneous gestures that speakers produce when they talk about a task reflect aspects of the speakers' knowledge about that task, aspects that are often not found in the speech that accompanies the gestures. But gesture can go beyond reflecting a speaker's current knowledge—it frequently presages the next steps the speaker will take in acquiring new knowledge, suggesting that gesture may play a role in cognitive change. To investigate this hypothesis, we explore the functions gesture serves with respect to both communication (the effects gesture has on listeners) and cognition (the effects gesture has on speakers themselves). We also explore the mechanisms that underlie the production of gesture, and we provide evidence that gesture has roots in speech, visuospatial thinking, and action. Gesturing is not merely hand waving, nor is it merely a window into the mind. It can affect how we think and reason and, as such, offers a useful tool to both learners and researchers.

Key Words: communication, embodied cognition, expert, learning, novice, speech, transitional knowledge, visuospatial thinking, working memory

Introduction

When people talk, they move their hands. These hand movements produced in conjunction with speech are called *gestures*. Like spoken language, gestures have the potential to reveal our thinking. But because gestures use a different representational format than speech does, they have the potential to reveal aspects of our thinking that are not evident in speech. Gesture thus offers a useful tool for learning about, and possibly changing, how we think.

We begin by situating gesture within the world of nonverbal behavior and highlighting why we think gesture, a nontraditional topic for a handbook on higher cognition, could contribute to our understanding of thinking and reasoning. We then take the first step in the argument that gesture plays a role in thinking by showing that gesture is not mere hand waving. It conveys substantive information

and, importantly, can reveal thoughts not found in the speech it accompanies. Gesture thus offers a unique window into the mind of the speaker. We then provide evidence that a speaker's spontaneous gestures often presage the next steps in the speaker's thinking and learning. Finally, we explore the purpose gesture serves (the functions of gesture) and the processes underlying its production (the mechanisms of gesture). We end with thoughts about how the study of gesture can continue to contribute to our understanding of the mind.

What Is Gesture?

In 1969, Ekman and Friesen outlined five categories of nonverbal behaviors produced during communication, thus framing the field of nonverbal communication. These behaviors vary according to where they are produced, how they relate to speech, and whether they are under conscious control.

Two behaviors—*emblems* and *illustrators*—are what people typically mean when they use the term *gesture*. Emblems are conventionalized movements of the hand that have word- or phrase-like meanings and can, in fact, substitute for words. Take, for example, the OK emblem (thumb and index finger touch and form a circle) or the thumbs-up emblem. Emblems share many properties with spoken words and with the signs in conventional sign languages of the deaf: They are consciously produced with the intent to communicate; they encode information arbitrarily; and they adhere to standards of well-formedness. Just as one can mispronounce a word or sign, it is possible to produce the wrong form of an emblem (imagine producing the OK emblem with the thumb and fourth finger—it just doesn't work). Because they are codified, emblems can stand on their own and, in fact, are often produced without speech.

In contrast, illustrators depend on speech for their meaning. Because they are always produced with speech, they take on the intentionality of speech. However, illustrators rarely come under conscious control, and they do not have a right or wrong form. Take, for example, a speaker who says that she ran upstairs while gesturing her trajectory with her hands; she can carve out her path using a pointing hand, an open palm, or any other hand shape. In general, illustrators convey information holistically and imaginatively and thus differ from speech, which conveys information componentially and categorically (Goldin-Meadow, 2003; McNeill, 1992). As a result, illustrators have the potential to reveal aspects of thinking not evident in speech. It is because illustrators are produced as part of an intentional communicative act, but are constructed at the moment of speaking, that they are of interest to us. They participate in communication, yet they are not part of a codified system.

We focus in this chapter on illustrators, called *gesticulation* by Kendon (1980) and plain old *gesture* by McNeill (1992), the term we use here. Thus, for the most part, we ignore emblems, as well as the three remaining nonverbal behaviors identified by Ekman and Friesen (1969): *Affect displays*, whose primary site is the face, convey the speaker's emotions, or at least those emotions that the speaker does not wish to mask (Ekman, Friesen, & Ellsworth, 1972). *Regulators*, which typically involve head movements or slight changes in body position, maintain the give and take between speaker and listener and help pace the exchange. *Self-adaptors*, which are fragments or reductions of previously learned adaptive hand movements,

are maintained by habit; for example, smoothing the hair, pushing glasses up the nose even when they are perfectly positioned, holding or rubbing the chin.

Gestures can mark the tempo of speech (beat gestures), point out referents of speech (deictic gestures), or exploit imagery to elaborate the contents of speech (iconic or metaphoric gestures). One question that is ripe for future research is whether these different types of gestures serve different functions and are served by different mechanisms. There is, in fact, evidence that beats and iconic/metaphoric gestures respond differently to the presence or absence of a listener (Alibali, Heath, & Myers, 2001), and that beat gestures are affected less by variation in the conceptual difficulty of speech than iconic or metaphoric gestures (Kita & Davies, 2009). Further research is needed to verify and explain these patterns. For now, we focus on deictic, iconic, and metaphoric gestures, as these are the gestures that have the potential to tell us the most about what a speaker is thinking.

Gesture Is Not Mindless Hand Waving and Often Reveals Thoughts Not Found in Speech

Gestures are interesting because they appear to provide a unique window onto thinking in that they reveal reliable information about a speaker's thoughts not evident in other behaviors. Accordingly, gesture can be a useful tool for exploring thinking and reasoning. As an example, gesture can reveal information about a speaker's prior motor experience that is not expressed in the accompanying speech. Cook and Tanenhaus (2009) asked adults to explain their solutions to the Tower of Hanoi problem after either solving the problem on the computer or solving it with real disks. The problem-solvers' verbal explanations were identical across the two groups (naïve observers could not distinguish the explanations produced by adults who had solved the problem on the computer from those produced by adults who had solved the problem using real disks). But their gestures differed. Adults who had solved the problem with real disks traced the trajectory of the disk with their hands (they mimed moving the disk up and over each peg). In contrast, adults who had solved the problem on the computer moved their hands laterally, mimicking the way the disks are moved on the screen (i.e., they do not have to be taken off the pegs before they are moved). The adults thus provided reliable cues about the problem-solving experiences they had had, cues that were not evident in their speech (Cook & Tanenhaus, 2009).

Gesture is not limited to displaying motor information and has been shown to reveal conceptual knowledge as well. Consider a child asked to participate in a series of Piagetian conservations tasks—the experimenter pours water from a tall, thin glass into a short, wide dish and asks the child whether the amount of water is still the same after the pouring. To succeed on this task and understand conservation of quantity, children need to integrate information across multiple dimensions—the height of the water in the container and its width. Nonconserving children focus on only one dimension, height or width, but not both. However, at times, a child will focus on one dimension in speech but provide evidence that he understands something about the importance of the second dimension in gesture. The child says that the amount of water is “different because this one is taller than that one,” thus focusing on the height of the water in speech. At the same time, he places a narrow C-shaped hand near the tall thin container, followed by a wider C-shaped hand near the short wide container, thus focusing on width in gesture—he displays knowledge of the second dimension *only* in his hands (Church & Goldin-Meadow, 1986). Note that children need to appreciate the compensatory relation between height and width in order to understand conservation of liquid quantity. Although this child appears to be firmly convinced that pouring the water alters its quantity, his hands reveal the first inkling that he may be ready to change his mind.

As a second example, gesture reveals knowledge that is relevant to understanding mathematical equivalence but is not evident in speech. The child is asked to solve problems like $3 + 4 + 6 = \underline{\quad} + 6$. To solve these problems correctly, children need to consider the relation between the two sides of the equation rather than simply adding up all of the numbers on the left side of the equation, or adding up all of the numbers in the problem (two common errors that children make when solving problems of this type). As in the conservation example, children sometimes produce gestures that reflect problem representations not expressed in the accompanying speech. For example, on the $3 + 4 + 6 = \underline{\quad} + 6$ problem, a child puts 19, an incorrect answer, in the blank and says, “I added the three, the four, the six, and the six to get nineteen” (an add-all-numbers strategy). At the same time, the child sweeps her left hand under the left side of the equation and then produces the same sweeping motion under the right side of the equation (an equalizer strategy), thus

displaying an awareness that the equation has two sides that should be treated alike (Alibali & Goldin-Meadow, 1993; Perry, Church, & Goldin-Meadow, 1988). Here again, the child displays an incorrect understanding of the problem in speech, but her hands reveal the first inkling that she may be ready to change her mind.

Gesture reveals aspects of children’s early cognitive development at a variety of ages and with respect to a variety of tasks. For example, toddlers reveal an understanding of one-to-one correspondence in the gestures they use in early counting before they display the same level of understanding in speech, and successful counting is associated with these gestural behaviors (Graham, 1999). Preschoolers (particularly boys) reveal an ability to mentally rotate shapes in their gestures not evident in their speech, and children whose gestures portrayed the spatial transformations were particularly successful at solving the mental transformation problems (Ehrlich, Levine, & Goldin-Meadow, 2006). Finally, early elementary school children solving balance problems reveal an understanding of the problems in their gestures that is not found in their speech (Pine, Lufkin, Kirk, & Messer, 2007), as do sixth grade children learning about plate tectonics (Singer, Radinsky, & Goldman, 2008) and preschool children learning to solve simple problems involving gears (Boncoddo, Dixon, & Kelley, 2010). Gesture can thus be used to make reliable inferences about children’s thinking across development.

Adults also reveal knowledge in their gestures that they do not display in their speech. For example, Alibali and colleagues (1999) asked adults first to describe algebra word problems about constant change, and then to indicate how they would go about solving the problems. The problems could be solved using either a discrete or continuous problem-solving strategy. Adults would often express one type of strategy in speech (e.g., continuous) while at the same time expressing the other type of strategy in gesture (discrete). Interestingly, speech and gesture taken together provided a more accurate picture of the strategy the adults planned to use to solve the problem than speech alone. Along the same lines, Garber and Goldin-Meadow (2002) found that speech and gesture taken together provided insight into the moments when adults (and children) were considering alternative routes in solving the Tower of Hanoi problem, moments that were not detectable from speech alone.

Finally, we see the same phenomenon in children at the early stages of language learning. For example,

when children begin to express causal relationships in speech, 3-year-olds use gesture to reinforce the goal of an action and 5-year-olds use gesture to add information about the instrument or direction of the action, information that is often not found in the accompanying speech; for example, producing an iconic *throw* gesture that adds information about the instrument to the utterance, “he broke the window” (Göksun, Hirsh-Pasek, & Golinkoff, 2010). As another example, when children begin to describe motion events in speech (e.g., “it went under there”), gesture is often used to reinforce or add information about manner, path, source, and endpoint. The type of information children choose to convey in gesture reflects not only their understanding of the event but also the linguistic framing of the particular language they are learning. For example, English allows speakers to combine manner and path within a single clause (“he rolls down”), and the gestures English speakers produce parallel this arrangement (the hand rolls as it moves down); in contrast, Turkish allows speakers to separate manner and path into separate clauses (the analog in English would be “he goes down by rolling”), and the gestures Turkish speakers produce reflect this structure (the hand rolls in place, followed by the hand moving down) (Özyürek, Kita, Allen, Furman, & Brown, 2005; Özyürek & Özçalışkan, 2000). Cross-linguistic studies of gesture can thus provide insight into how children come to describe events in the manner typical of their language.

Gestures Presage Next Steps in Thinking and Learning

Gestures not only reveal a person’s thinking at the time that they are produced, but they also forecast subsequent changes in thinking. Gesture has been found to reliably predict future thinking across a wide variety of domains. In fact, the data suggest that gesture is often a more useful predictor of subsequent thinking than the concurrent speech. We begin by examining gesture’s ability to foreshadow changes in child language learning.

Learning Language

Children’s early gestures have been shown to foreshadow their subsequent vocabulary development (Bavin et al., 2008; Goodwyn & Acredolo, 1993; Rowe & Goldin-Meadow, 2008, 2009; Rowe, Özçalışkan, & Goldin-Meadow, 2008). For example, a child’s early deictic gestures reliably predict which nouns are likely to enter that child’s

spoken vocabulary in the next 3 months (Iverson & Goldin-Meadow, 2005).

Early gesture not only predicts the particular words children are likely to learn but also when and how those words are combined with one another. A child’s early single-word utterances are often accompanied by gesture, and the relation between these early gestures and the speech they accompany reliably predicts when the child will produce her first two-word utterance. Children whose gestures overlap in meaning with the accompanying speech (e.g., pointing at a cup while saying “cup”) are likely to remain in the single-word stage for many months. In contrast, children whose gestures convey a different meaning from the accompanying speech (e.g., pointing at a cup while saying “mine”) are likely to begin combining words into two-word combinations within the next few months (Goldin-Meadow & Butcher, 2003; Iverson & Goldin-Meadow, 2005). In fact, the particular constructions expressed in gesture + speech combinations can be used to predict the emergence of the same constructions in speech later in development. For example, a child who conveys an action predicate plus an object argument in speech and gesture (e.g., “open” combined with a point at a box) is likely to produce an action predicate + object argument construction entirely in speech (“open box”) several months later (Özçalışkan & Goldin-Meadow, 2005).

Gesture continues to forecast children’s verbal milestones beyond the transition from one-word to two-word speech. For example, children produce their first complex sentence containing two predicates in gesture and speech (e.g., “I like it,” said while producing an *eat* gesture) several months before producing their first complex sentence entirely in speech (“I like eating it;” Özçalışkan & Goldin-Meadow, 2005). Interestingly, although children rely on gesture to produce the first instance of a construction (e.g., a predicate plus one argument, “give” + point at cookie), once the construction is established in their repertoire, children are no more likely to use gesture to flesh out the construction than they are to use speech. For example, they are just as likely to produce their first predicate plus three arguments entirely in speech (“*you see my butterfly on my wall*”) as they are to produce their first predicate plus three arguments in gesture and speech (“*Daddy clean all the bird poopie*” + point at *table*) (Özçalışkan & Goldin-Meadow, 2009). Gesture thus acts as a harbinger of linguistic steps only when those steps involve new constructions, not when the steps merely flesh out existing constructions.

As these findings suggest, gesture is not a global index of subsequent linguistic change but rather an indication of specific changes. Rowe and Goldin-Meadow (2009) observed 52 children interacting with their caregivers at home and found that gesture use at 18 months *selectively* predicted lexical versus syntactic skills at 42 months, even with early child speech controlled. Specifically, the number of different meanings children conveyed in gesture at 18 months predicted the size of their spoken vocabularies at 42 months, but the number of gesture + speech combinations did not. In contrast, the number of gesture + speech combinations, particularly those conveying sentence-like ideas, produced at 18 months predicted sentence complexity at 42 months, but meanings conveyed in gesture did not. Particular milestones in vocabulary and sentence complexity at age 3 1/2 years can thus be predicted from the way children moved their hands 2 years earlier.

Importantly, not only does gesture predict language development in typically developing children, but it also predicts subsequent language development in atypical populations. For example, some children who are late talkers will “catch up” to their typically developing peers, whereas others will continue to have persistent delays in language production. The interesting result is that early gesture can predict which children will catch up and which children will not (Thal, Tobias, & Morrison, 1991; Thal & Tobias, 1992); the children who caught up performed well on two gesture tasks: They could imitate object-related gestures produced by the experimenter (e.g., making a toy airplane fly), and they could reproduce a series of familiar, scripted actions modeled by an experimenter (e.g., feeding a teddy bear by putting him in a highchair, putting on his bib, feeding him an apple, and wiping his mouth). Gesture can also predict which children with early unilateral focal brain injury are likely to remain delayed with respect to vocabulary development, and which children are likely to move into the normal range. Children with brain injury who produced a repertoire of gestures at 18 months comparable to the repertoire of gestures produced by typically developing 18-month-old children were subsequently within the normal range of spoken vocabulary development at 22, 26, and 30 months. In contrast, children with brain injury whose gesture production at 18 months was outside of the typical range continued to show delays in vocabulary development at 22, 26, and 30 months (Sauer, Levine, & Goldin-Meadow, 2010). As a final example, early gesture

appears to be a more robust predictor of subsequent language development in children with autism than other social communication factors (Luyster, Kedlec, Carter, & Tager-Flusberg, 2008; see also Smith, Mirenda, & Zaidman-Zait, 2007). Gesture is thus an early marker that can be used to determine whether children whose language-learning trajectory has the potential to go astray will, in fact, experience delay. In this sense, gesture is a promising tool for diagnosing persistent delay.

Learning Other Cognitive Tasks

Children enter language learning hands first. But they continue to gesture even after having mastered the rudiments of language. At that point, children’s gestures begin to forecast changes in their thinking in other areas of cognitive development. One important experimental difference between the studies of gesture in learning language versus learning other cognitive tasks is that the language studies are all longitudinal observations of children in naturalistic settings. We see variability in the gestures children spontaneously produce at an early time point, and we use that variability to predict the onset of linguistic constructions at a later time point. We assume that the early gesture producers are ready to learn these linguistic constructions and need only more time or more input to do so.

In contrast, the studies of children learning other cognitive tasks tend to be short-term experimental manipulations. We again see variability in the gestures children spontaneously produce, this time with respect to a particular task, say, conservation of liquid quantity. But rather than wait for the children to experience additional input, we give the children instruction in the task and observe which children profit from that instruction. Recall the child described earlier who talked about the height of the water in speech but indicated its width in gesture. Although this child says that the amount of water is different when it is poured from one container to another (i.e., he is a nonconserver), his gestures indicate that he knows more about the task than his words indicate. And, indeed, when given instruction in conservation, this child is likely to make progress on the task—more likely than a child who focuses on the height of the water in both speech and gesture (Church & Goldin-Meadow, 1986).

As another example, consider the child described earlier who was asked how she arrived at her incorrect answer to a mathematical equivalence problem and produced an add-all-numbers strategy in speech while

at the same time producing an equalizer strategy in gesture. Here again, the child's gestures indicate that she knows more about mathematical equivalence than her words indicate. When given instruction in the problem, the child is likely to profit from that instruction and learn how to solve problems of this type correctly—more likely than a child who gives an add-all-numbers strategy in both speech and gesture (Perry, Church, & Goldin-Meadow, 1988). Moreover, when children's responses are charted during the course of instruction, we can see a child systematically progress through three periods characterized by the relation between gesture and speech—the child produces (1) the same strategy in both speech and gesture and that strategy is incorrect (e.g., add-all-numbers); (2) two different strategies, one in speech (e.g., add-all-numbers) and a different one in gesture (e.g., equalizer); (3) the same strategy in both speech and gesture but now the strategy is correct (e.g., equalizer) (Alibali & Goldin-Meadow, 1993). If, as in this case, only one of the modalities conveys a correct strategy, that correct strategy is often found in gesture rather than speech. Gesture, when taken in relation to speech, signals that the child is ready to take the next step in learning about mathematical equivalence. Interestingly, when a child fails to pass through step (2) and goes directly from step (1) to step (3), the child's understanding of mathematical equivalence is relatively fragile; in particular, the child is unable to generalize the knowledge gained during instruction and does not retain the knowledge on a follow-up test (Alibali & Goldin-Meadow, 1993).

We see this phenomenon on a variety of tasks and ages. For example, as described earlier, elementary school children asked to reason about balance often express new ideas about the task in gesture before expressing these same ideas in speech (Pine et al., 2007). When given instruction in the task, these children are the ones most likely to benefit from that instruction (Pine, Lufkin, & Messer, 2004). A similar effect has been found in adult learners asked to predict which way the last gear in a configuration of gears will turn (Perry & Elder, 1997) or asked to draw the stereoisomer of a molecule (Larson et al., 2010). In both cases, adults who display task-relevant information in their gestures not found in their speech are particularly likely to make progress on the task after getting instruction in the task.

It is clear that gesture offers a window onto thinking, and that the picture provided by gesture is often different from the view provided by speech. But why does gesture offer this privileged view? We

explore first the functions gesture serves and then the mechanisms underlying its production to better understand why and how gesture precedes and predicts our thinking and reasoning.

The Functions Gesture Serves: What Does Gesture Do? *Communication: The Impact of Gesture on the Listener*

We now know that speakers' gestures reveal their thoughts. Accordingly, one function that gesture could serve is to convey those thoughts to listeners. For gesture to serve this function, listeners must be able to extract information from the gestures they see. And, indeed, there is considerable evidence that listeners can use gesture as a source of information about the speaker's thinking (e.g., Graham & Argyle, 1975; McNeil, Alibali, & Evans, 2000).

The ability of listeners to glean information from a speaker's gestures can be seen most clearly when the gestures convey information that cannot be found anywhere in the speaker's words. Take, for example, the Cook and Tanenhaus (2009) Tower of Hanoi study described earlier in which speakers conveyed information about the trajectory a disk followed as it was moved from one peg to another—either an arced trajectory that went up and over the peg, or a lateral trajectory that ignored the peg. This information was *not* represented in the speakers' words. Listeners who saw the arced gestures were more likely to move the disk up and over the peg when they were later asked to solve the Tower of Hanoi problem on the computer (where it is not necessary to arc the disks to move them) than listeners who saw the lateral gestures (Cook & Tanenhaus, 2009). The listeners had not only read the action information off of the speakers' gestures, but that information had had an effect on their own subsequent actions.

Adults can also glean information from child speakers. When adult listeners are asked to describe the responses child speakers give on a conservation task, the adults frequently describe information that the children expressed *only* in gesture and not in speech (Goldin-Meadow & Momeni Sandhofer, 1999), making it clear that listeners can glean substantive information from speakers' gestures.

Perhaps the clearest example of this phenomenon is when the listener translates the information conveyed in the speaker's gestures into speech. Take, for example, a listener retelling a story in which the speaker said, "She whacks him one," while producing a punching gesture. The listener subsequently

redescribed this event as “She punches Sylvester out” (Cassell, McNeill, & McCullough, 1999); she had not only seen and interpreted the speaker’s punching gesture but also integrated the information into her speech (see also Goldin-Meadow, Kim, & Singer, 1999; Goldin-Meadow & Singer, 2003). Similarly, mothers of young language-learning children frequently respond to their children’s early gestures by translating them into speech (e.g., saying, “Yes, the bird is napping,” in response to a child’s point at a bird produced while saying “nap;” Goldin-Meadow, Goodrich, Sauer, & Iverson, 2007).

Not surprisingly, listeners increase their reliance on the speaker’s gestures in situations where speech is difficult to understand; for example, when there is noise in the speech signal (Holle, Obleser, Rueschemeyer, & Gunter, 2010; Rogers, 1978; Thompson & Massaro, 1986, 1994). Listeners are also particularly influenced by gesture when the spoken message is relatively complex (McNeil, Alibali, & Evans, 2000).

Gesture can even affect the information listeners glean from the accompanying speech. Listeners are faster to identify a speaker’s referent when speech is accompanied by gesture than when it is not (Silverman, Bennetto, Campana, & Tanenhaus, 2010). When processing speech that is accompanied by gesture conveying the same information, listeners are *more* likely to glean the message from speech than when processing speech accompanied by no gesture (Beattie & Shovelton, 1999, 2002; Graham & Argyle, 1975; McNeil et al., 2000; Thompson & Massaro, 1994). Conversely, when processing speech that is accompanied by gesture conveying different information, listeners are *less* likely to glean the message from speech than when processing speech accompanied by no gesture (Goldin-Meadow & Momeni Sandhofer, 1999; Kelly & Church, 1998; McNeil et al., 2000). In addition, more incongruent gestures lead to greater processing difficulty than congruent gestures (Kelly, Özyürek, & Maris, 2010). The effect that gesture has on listeners’ processing is thus linked to the meaning relation between gesture and speech. Moreover, listeners cannot ignore gesture even when given explicit instructions to do so (Kelly, Özyürek, & Maris, 2010; Langton, O’Malley, & Bruce, 1996), suggesting that the integration of gesture and speech is automatic.

Like adults, children are able to extract information from a speaker’s gestures, even when the information is not conveyed in the accompanying speech

(Kelly & Church, 1997). Very young children can use gesture as a source of information to support word learning (Booth, McGregor, & Rohlfing, 2008; McGregor, Rohlfing, Bean, & Marschner, 2008). By age 3 years, children are able to integrate information across speech and gesture (Kelly, 2001; Morford & Goldin-Meadow, 1992; Thompson & Massaro, 1986). However, the influence that gesture has on how speech is interpreted does appear to increase throughout childhood (Thompson & Massaro, 1986, 1994).

The fact that gesture can communicate information to listeners suggests that gesture might be particularly helpful in teaching and learning situations. Indeed, child listeners have been shown to learn more from a lesson that contains gesture than from a lesson that does not contain gesture (Church, Ayman-Nolley, & Mahootian, 2004; Valenzano, Alibali, & Klatzky, 2003), even when the gestures are not directed at objects in the immediate environment (Ping & Goldin-Meadow, 2008). Interestingly, even though communication often suffers when speakers produce gestures that convey different information from their speech (Goldin-Meadow & Momeni Sandhofer, 1999; Kelly & Church, 1998; McNeil et al., 2000), children learning mathematical equivalence seem to benefit most from instruction that contains one correct strategy in speech and a different correct strategy in gesture (Singer & Goldin-Meadow, 2005); that is, from instruction in which gesture conveys different information from speech. One possibility is that, in these instances, the additional information in gesture makes it more likely that one of the representations in the instruction matches the child’s next developmental state and, in this way, facilitates learning.

The process by which gesture affects the listener is currently being explored using a variety of brain imaging paradigms. Using functional magnetic resonance imaging (fMRI), researchers have found that gesture activates areas associated with language processing, including Broca’s area (Skipper, Goldin-Meadow, Nusbaum, & Small, 2007; Willems, Özyürek, & Hagoort, 2007). Gesture also appears to affect how processing is organized by influencing the connectivity among the relevant brain regions (Skipper, Goldin-Meadow, Nusbaum, & Small, 2007).

Using electroencephalography (EEG), a number of researchers have demonstrated that the relation between gesture and speech can modulate brain activity. Gestures that are semantically anomalous with respect to the accompanying speech are

associated with a more negative N400 waveform (Bernardis, Salillas, & Caramelli, 2008; Holle & Gunter, 2007; Kelly, Kravitz, & Hopkins, 2004; Özyürek Willems, Kita, & Hagoort, 2007; Wu & Coulson, 2005, 2007); the N400 is known to be sensitive to incongruent semantic information (Kutas & Hillyard, 1984). For example, gestures conveying information that is truly incongruent with the information conveyed in speech (gesturing *short* while saying “tall”) produce a large negativity at 400 ms (Kelly et al., 2004). Interestingly, gestures conveying information that is different from, but complementary to, information conveyed in speech (gesturing *thin* while saying “tall” to describe a tall, thin container) are processed no differently at this stage from gestures that convey the same information as speech (gesturing *tall* while saying “tall;” Kelly et al., 2004). Neither one produces a large negativity at 400 ms; that is, neither one is recognized as a semantic anomaly. It is important to note, however, that at early stages of sensory/phonological processing (P1-N1 and P2), speech accompanied by gestures conveying different but complementary information (e.g., gesturing *thin* while saying “tall”) is processed differently from speech accompanied by gestures conveying the same information (gesturing *tall* while saying “tall”). Thus, complementary differences between the modalities (i.e., the information conveyed in gesture is different from, but has the potential to be integrated with, the information conveyed in speech) are noted at early stages of processing, but not at later, higher level stages.

Gestures can affect the message listeners glean from speakers. Nonetheless, it is not clear that speakers *intend* their gestures to be communicative. Some gestures are meant to be communicative; for example, gestures that are referred to explicitly in the accompanying speech (“this one,” accompanied by a pointing gesture). However, it is not clear whether gestures that are not explicitly referenced in speech are intended to be communicative. One way to explore this issue is to vary whether speakers and listeners have visual access to one another. The question is whether speakers will gesture even when their listeners cannot see them and thus cannot acquire any information from those gestures. The answer is that speakers gesture less frequently when their listeners do not have visual access to gesture, particularly iconic gestures (Alibali et al., 2001; Cohen, 1977; Emmorey & Casey, 2001). However, speakers do not stop gesturing completely when their listeners cannot see them (Alibali et al., 2001;

Bavelas, Chovil, Lawrie, & Wade, 1992; Bavelas, Gerwing, Sutton, & Prevost, 2008; Cohen, 1977; Cohen & Harrison, 1973; Emmorey & Casey, 2001; Krauss, Dushay, Chen, & Rauscher, 1995; Rimé, 1982), suggesting that gesture may be produced for the benefit of the speaker as well as the listener. The next section explores the functions gesture can serve for the speaker.

Cognition: The Impact of Gesture on the Speaker

FACILITATING LEXICAL ACCESS

Gestures have long been argued to help speakers “find” words, that is, to facilitate lexical access (Rauscher, Krauss, & Chen, 1996). Consistent with this hypothesis, speakers are more likely to gesture when they are producing unrehearsed speech (Chawla & Krauss, 1994), when they are about to produce less predictable words (Beattie & Shovelton, 2000), and when lexical access is made more difficult (Rauscher, Krauss, & Chen, 1996). Temporally, gestures precede less familiar words to a greater degree than they precede more familiar words (Morrel-Samuels & Krauss, 1992). And brain-damaged patients with difficulties in lexical access (that is, patients with aphasia) gesture at a higher rate than patients with visuospatial deficits (Hadar, Burstein, Krauss, & Soroker, 1998). These findings suggest that gesture is associated with difficulties in lexical access. More direct evidence that gesture plays a role in lexical access comes from reports that speakers are more successful at resolving tip-of-the-tongue states when they are permitted to gesture than when they are not, for both adult (Frick-Horbury & Guttentag, 1998) and child (Pine, Bird, & Kirk, 2007) speakers (but see Beattie & Coughlan, 1999).

REDUCING DEMANDS ON CONCEPTUALIZATION

Speakers gesture more on problems that are conceptually difficult, even when lexical demands are equated (Alibali, Kita, & Young, 2000; Hostetter, Alibali, & Kita, 2007; Kita & Davies, 2009; Melinger & Kita, 2007). As an example, when adult speakers are asked to describe dot patterns, they gesture more when talking about patterns that do not have lines connecting the dots (patterns that are more difficult to conceptualize) than patterns that do have lines (Hostetter et al., 2007). As a second example, children who are asked to solve Piagetian conservation problems (problems that require conceptualization) gesture more than when they are

simply asked to describe the materials used in the conservation problems (Alibali et al., 2000).

Gesture may be particularly effective in reducing conceptual demands in visuospatial tasks, as gesture is a natural format for capturing spatial information. Gesture has, in fact, been found to facilitate visuospatial processing in speakers, either by maintaining visuospatial information in memory (Morsella & Krauss, 2004; Wesp, Hesse, Keutmann, & Wheaton, 2001) or by facilitating packaging of visuospatial information for spoken language (Kita, 2000). Gesture can also facilitate transformation of spatial information in memory; when performing mental rotation tasks, adults are particularly successful if they produce gestures (Chu & Kita, 2008; Schwartz & Black, 1999) or hand movements (Wexler, Kosslyn, & Berthoz, 1998; Wohlschlager & Wohlschlager, 1998) consistent with the actual rotation that is to be performed, or consistent with the movement that would activate the rotation (e.g., a pulling gesture that mimics pulling a string from a spool to make the spool turn; Schwartz & Holton, 2000).

Although these findings are consistent with the idea that gesturing reduces demands on conceptualization, the relevant studies manipulated conceptualization difficulty and observed the effects of the manipulation on gesturing, finding that conceptualization difficulty and gesturing go hand in hand. But to be certain that gesturing plays a role in reducing conceptualization demands (as opposed to merely reflecting those demands), future work will need to manipulate gesture and demonstrate that the manipulation reduces the demands on conceptualization.

REDUCING DEMANDS ON WORKING MEMORY

Gesturing has been shown to reduce demand on speakers' working memory. When asked to remember an unrelated list of items while explaining how they solved a math problem, speakers are able to maintain more items in verbal working memory (and thus recall more items) when they gesture during the explanation than when they do not gesture. This effect has been found in both children and adults (Goldin-Meadow, Nusbaum, Kelly, & Wagner, 2001). Interestingly, the effect has also been found for items in visual working memory (i.e., speakers maintain more items in visual working memory when they gesture during their explanations than when they do not gesture; Wagner, Nusbaum, & Goldin-Meadow, 2004), suggesting that gesturing lightens the load on working memory whether the

stored items are visual or verbal. In addition, gesturing reduces demand on working memory even when the gestures are not directed at visually present objects (Ping & Goldin-Meadow, 2010), suggesting that gesturing confers its benefits by more than simply tying abstract speech to objects directly visible in the environment.

Importantly, it is not just moving the hands that reduces demand on working memory—it is the fact that the moving hands convey meaning. Producing gestures that convey different information from speech reduces demand on working memory *less* than producing gestures that convey the same information in speakers who are experts on the task (Wagner et al., 2004). Interestingly, we find the opposite effect in speakers who are novices—producing gestures that convey different information from speech reduces demand on working memory *more* than producing gestures that convey the same information as speech (Ping & Goldin-Meadow, 2010). In both cases, however, it is the meaning relation that gesture holds to speech that determines, at least in part, the extent to which the load on working memory is reduced.

LINKING INTERNAL REPRESENTATIONS TO THE WORLD

Gesturing may help link the speaker's internal representations to the physical and communicative environment. Deictic gestures, in particular, may facilitate speakers' use of the surrounding space (Ballard, Hayhoe, Pook, & Rao, 1997). For example, for children learning to count, gesture seems to be important in coordinating number words with objects and in keeping track of which objects have already been counted (Saxe & Kaplan, 1981). Alibali and DiRusso (1999) explored gesture's role in children's counting by comparing three conditions: the child gestured while counting, the child was restricted from gesturing while counting, and the child watched a puppet gesture while the child counted. They found that children were most accurate when their counting was accompanied by gesture, theirs or the puppet's. But they were least likely to make errors coordinating number words and objects when the children themselves produced the gestures.

However, as mentioned earlier, gestures do not have to be directed at visible objects in order for speakers to benefit from gesturing. Ping and Goldin-Meadow (2010) measured demand on working memory in children asked to remember

an unrelated list of items while explaining their responses to a conservation task. Children were told to gesture on half the trials and not to gesture on the other half. One group gave their explanations with the task objects present; the other group gave their explanations with the task objects out of view. Children remembered more items, reflecting a reduced demand on working memory, when they gestured during their explanations than when they did not gesture, even when the objects were not visible. Gesturing does not need to be tied to the physical environment in order to be effective. Indeed, over the course of learning a task, gestures can become more and more removed from the immediate physical environment, eventually becoming internalized (Chu & Kita, 2008).

ACTIVATING OLD KNOWLEDGE AND BRINGING IN NEW KNOWLEDGE

Gesturing can activate knowledge that the speaker has but does not express. Broaders, Cook, Mitchell, and Goldin-Meadow (2007) asked children to explain how they solved six mathematical equivalence problems with no instructions about what to do with their hands. They then asked the children to solve a second set of comparable problems and divided the children into three groups: Some were told to move their hands as they explained their solutions to this second set of problems; some were told not to move their hands; and some were given no instructions about their hands. Children who were told to gesture on the second set of problems added strategies to their repertoires that they had not previously produced; children who were told not to gesture and children given no instructions at all did not. Most of the added strategies were produced in gesture and not in speech and, surprisingly, most were correct. In addition, when later given instruction in mathematical equivalence, it was the children who had been told to gesture, and who had added strategies to their repertoires, who subsequently profited from the instruction and learned how to solve the math problems. Being told to gesture thus encouraged children to express ideas that they had previously not expressed, which, in turn, led to learning.

But can gesture, on its own, create new ideas? To determine whether gesture can create new ideas, we need to teach speakers to move their hands in particular ways. If speakers can extract meaning from their hand movements, they should be sensitive to the particular movements they are taught to produce

and learn accordingly. Alternatively, all that may matter is that speakers move their hands. If so, they should learn regardless of which movements they produce. To investigate these alternatives, Goldin-Meadow, Cook, and Mitchell (2009) manipulated gesturing during a math lesson. They found that children required to produce *correct* gestures learned more than children required to produce *partially correct* gestures, who learned more than children required to produce *no* gestures. This effect was mediated by whether, after the lesson, the children added information to their spoken repertoire that they had conveyed only in their gestures during the lesson (and that the teacher had not conveyed at all). The findings suggest that gesture is involved not only in processing old ideas but also in creating new ones. We may be able to lay the foundations for new knowledge simply by telling learners how to move their hands (see Cook, Mitchell & Goldin-Meadow, 2008 for related findings) or by moving our hands ourselves (children who see their teachers gesture a concept are likely to gesture themselves and, in turn, are likely to learn the concept; Cook & Goldin-Meadow, 2006).

The Mechanism Underlying Gesture Production: Where Does Gesture Come From?

Gesture is not simply mindless hand waving. It offers a window onto speakers' thinking, affording access to information not available in the speakers' other behaviors. But gesture does more than simply externalize speakers' thinking. When speakers gesture, those gestures have an impact not only on their listeners but also on their own cognition. We next explore the mechanism that underlies gesture production.

Roots in Speech

Gestures are produced in conjunction with speech. One mechanism that could underlie the production of gesture is speech production; that is, the processes supporting speech production may naturally lead to gesture production.

It is clear that gesture and speech are inexorably linked. Congenitally blind speakers, who have never seen another person gesture, produce gestures when they speak, even when speaking to blind listeners (Iverson & Goldin-Meadow, 1997, 1998). Prior to speaking, children produce rhythmic hand movements in conjunction with their vocal babbling (Masataka, 2001). Although gestures are sometimes

produced without accompanying speech, the vast majority of gestures are produced while speaking (McNeill, 1992), suggesting that speech and gesture production may share a single mechanism. Moreover, even when speakers do not produce overt gestures, recalling concrete and spatial words from definitions is associated with changes in muscle potentials in the arms (Morsella & Krauss, 2005). More generally, speaking is associated with increases in corticospi-
nal excitability of hand motor areas (Meister et al., 2003; Seyal, Mull, Bhullar, Ahmad, & Gage, 1999; Tokimura, Tokimura, Oliviero, Asakura, & Rothwell, 1996). Listening to speech has also been associated with activity in the hand motor cortex (Flöel, Ellger, Breitenstein, & Knecht, 2003). Production of speech and production of hand movements are thus tightly linked to one another, at both the behavioral and the neural level.

Gesture is linked to spoken language at every level of analysis, including the phonological level, lexical level, syntactic level, prosodic level, and conceptual level (as discussed earlier in the section on “The Functions Gesture Serves”). At the phonological level, producing hand gestures influences the voice spectra of the accompanying speech for deictic gestures (Chieffi, Secchi, & Gentilucci, 2009), emblem gestures (Barbieri, Buonocore, Dalla Volta, & Gentilucci, 2009; Bernardis & Gentilucci, 2006), and beat gestures (Krahmer & Swerts, 2007). When phonological production breaks down, as in stuttering or aphasia, gesture production stops as well (Mayberry & Jacques, 2000; McNeill, Levy, & Pedelty, 1990). There are phonological costs to producing gestures with speech—producing words and deictic gestures together leads to long initiation times for the accompanying speech, relative to producing speech alone (Feyereisen, 1997; Levelt, Richardson, & Laheij, 1985). Viewing gesture also affects voicing in listeners’ vocal responses to audio-visual stimuli (Bernardis & Gentilucci, 2006).

At the lexical level, as discussed earlier, gesturing increases when the speaker is searching for a word. More generally, gestures both reflect, and compensate for, gaps in a speaker’s verbal lexicon. Gestures can package information in the same way that information is packaged in the lexicon of the speaker’s language. For example, when speakers of English, Japanese, and Turkish are asked to describe a scene in which an animated figure swings on a rope, English speakers overwhelmingly use the verb “swing” along with an arced gesture (Kita & Özyürek, 2003). In contrast, speakers of Japanese

and Turkish, languages that do not have single verbs that express an arced trajectory, use generic motion verbs along with the comparable gesture; that is, a straight gesture (Kita & Özyürek, 2003). But gesture can also compensate for gaps in the speaker’s lexicon by conveying information that is not encoded in the accompanying speech. For example, complex shapes that are difficult to describe in speech can be conveyed in gesture (Emmorey & Casey, 2001).

At the syntactic level, as described earlier, gestures are influenced by the structural properties of the accompanying speech. For example, English expresses manner and path within the same clause, whereas Turkish expresses the two in separate clauses. The gestures that accompany manner and path constructions in these two languages display a parallel structure—English speakers produce a single gesture combining manner and path (a rolling movement produced while moving the hand forward), whereas Turkish speakers produce two separate gestures (a rolling movement produced in place, followed by a moving forward movement) (Kita & Özyürek, 2003; Kita et al., 2007). Gesture production also reflects the amount of information encoded in a syntactic structure. Speakers gesture more when producing an unexpected (and, in this sense, more informative) syntactic structure than when producing an expected structure (Cook, Jaeger, & Tanenhaus, 2009).

At the prosodic level, the movement phase of a speaker’s gesture co-occurs with the point of peak prosodic emphasis in the accompanying clause in speech (Kendon, 1980; McClave, 1998). And listeners make inferences about the perceived prominence of words in an utterance from the speaker’s gestures (Krahmer & Swerts, 2007).

Gestures have also been linked to the conceptualization process involved in speaking, that is, the process by which speakers determine which information to linguistically encode in an utterance. Support for this hypothesis comes from studies showing that difficulty, or greater ambiguity about what to say, is associated with increases in gesture production (Alibali et al., 2000; Hostetter, Alibali, & Kita, 2007; Kita & Davies, 2009; Melinger & Kita, 2007). However, not all studies find that increases in conceptualization difficulty are associated with increases in gesture rate (Sassenberg & van der Meer, 2010).

An explanation on the evolutionary timespan for the close relationship between gesture and speech is that spoken language may have evolved from

more primitive gestural communication systems (Corballis, 1992; Fitch, 2000; Holden, 2004). If so, modern-day gestures that are produced in conjunction with speech may represent vestigial activity of a prior system for communication. Gesture may continue to contribute functionally to spoken communication, as the findings in the previous section suggest, or may simply reflect the underlying organization of the system without being functionally involved in spoken language production.

Roots in Visuospatial Thinking

Gestures could also emerge from visuospatial thinking. Consistent with this hypothesis, speakers are likely to gesture when talking about things that are spatial or imageable (Alibali et al., 2001; Beattie & Shovelton, 2002; Krauss, 1998; Lavergne & Kimura, 1987; Rauscher et al., 1996; Sousa-Poza, Rohrberg, & Mercure, 1979) and when conveying information that has been acquired visually (as opposed to verbally, Hostetter & Hopkins, 2002). In addition, when speakers are restricted from gesturing, the spatial (Graham & Heywood, 1975) and/or imagistic (Rime, Schiaratura, Hupet, & Ghysselinckx, 1984) content of the accompanying speech changes. Finally, brain-damaged patients with visuospatial deficits gesture less than comparable patients with lexical access deficits (Hadar et al., 1998).

In a striking example of the link between visuospatial representation and gesture production, Haviland (1993) described how a speaker of Guugu Yimithirr, an Australian language that uses an absolute rather than a relative reference frame to represent direction, adjusted his gesture production across two tellings of the same story so that his gestures were true to the actual spatial layout of the original event (i.e., his gestures were also absolute rather than relative). The speaker, describing how a boat overturned many years ago, produced a rolling motion away from his body when facing west since the boat had actually rolled from east to west. However, when telling the story on another occasion, he happened to be facing north rather than west. In this retelling, he produced the same rolling-over gesture, but this time his hands rolled from right to left rather than away from his body. The gesture was accurate with respect to the actual event (an east-to-west roll). Importantly, the speaker did not refer to the absolute spatial context of the original event in his speech.

Gestures are particularly likely to represent visuospatial thinking that involves transformations.

For example, gestures frequently represent orientations and rotations of block locations (Emmorey & Casey, 2001), spatial transformations (Trafton et al., 2006), and component motions of entities in physics problems (Hegarty, Mayer, Kriz, & Keehner, 2005; see Hegarty & Stull, Chapter 31).

In addition to representing spatial information directly, gestures may also reflect metaphoric use of visuospatial representations. Speakers use a wide variety of spatial metaphors when representing nonspatial concepts, including time (Alverson, 1994; Clark, 1973), mathematics (Lakoff & Nunéz, 2000), and emotions (Lakoff & Johnson, 1999). Moreover, these metaphoric representations are not simply linguistic conventions but can have an effect on information processing. For example, people's judgments about the meaning of a temporal expression like "the meeting has been moved forward 2 days" depends on how they envision themselves moving through space (Boroditsky & Ramscar, 2002). Gestures that are produced when talking about time are consistent with the underlying metaphoric mapping between space and time in the speaker's language (e.g., English speakers gesture in front of themselves when talking about the future, whereas speakers of Aymara gesture to their backs; Núñez & Sweetser, 2006). Producing appropriate hand movements can also facilitate comprehension of metaphoric expressions like "push the argument" (Wilson & Gibbs, 2007). Thus, gestures may reflect and engage visuospatial thinking when it is used metaphorically as well as when it is used literally.

Roots in Action

Hostetter and Alibali (2008) have proposed that gestures emerge from perceptual and motor simulations underlying the speaker's thoughts (see also Rimè & Schiaratura, 1991). This proposal is based on recent theories claiming that linguistic meaning is grounded in perceptual and action experiences (Barsalou, 1999; Glenberg & Kaschak, 2002; Richardson, Spivey, Barsalou, & McRae, 2003; Zwaan, Stanfield, & Yaxley, 2002). If so, gesture could be a natural outgrowth of the perceptual-motor experiences that underlie language. Under this view, the richer the simulations of the experiences, the more speech will be accompanied by gesture.

Support for the hypothesis that gestures emerge out of motor processes comes from a study conducted by Feyereisen and Havard (1999) who explored whether certain types of imagery, including motor imagery, are likely to lead to gesture. Speakers

were asked to describe motor activities (e.g., changing a tire, wrapping a present), visual scenes (e.g., rooms, landscapes), or abstract topics (e.g., women in politics, the death penalty). Motor imagery frequently resulted in iconic gestures, whereas abstract topics led to beat gestures. However, the weakness of the study is that topic and type of imagery were confounded. The action words generated to describe the topic, rather than motor imagery, might therefore have led to the frequent iconic gestures.

The study conducted by Cook and Tanenhaus (2009) that was described in an earlier section also explored the relation between motor imagery and gesture. In this study, speakers who performed the Tower of Hanoi task with real disks gestured differently from speakers who performed the task on the computer, even though their verbal descriptions were identical. The speakers who solved the problem with real disks produced gestures that simulated the actions they used to move the disks (i.e., they lifted the disk up and over the peg), suggesting that gestures can reflect motor representations. As another example, speakers gesture more when describing dot patterns that they constructed with wooden pieces than dot patterns that they viewed on a computer screen (Hostetter & Alibali, 2010).

But gestures do not merely reflect the action simulations that underlie the speaker's thinking; they can also influence which action components become part of the speaker's mental representation. Beilock and Goldin-Meadow (2010) asked adults to first solve the Tower of Hanoi problem with real, weighted disks (TOH1). The smallest disk in the tower was the lightest and could be lifted with one hand; the biggest was so heavy that it required two hands. The adults were then asked to explain how they solved the problem, gesturing while doing so. After the explanation, they solved the problem a second time (TOH2). For some problem solvers (*No-Switch Group*), the disks in TOH2 were identical to TOH1, and they, not surprisingly, improved on the task (they solved TOH2 in fewer moves and in less time than TOH1). For others (*Switch Group*), the disk weights in TOH2 were reversed—the smallest disk was now the heaviest and could no longer be lifted with one hand. This group did not improve and, in fact, took more moves and more time to solve the problem on TOH2 than TOH1. Importantly, however, the performance on the *Switch group* on TOH2 could be traced back to the gestures they produced during the explanation task: The more they used one-handed gestures when talking about

moving the smallest disk during the explanation, the worse they did on TOH2 (remember that the smallest disk on TOH2 in the *Switch group* could no longer be lifted with one hand). There was no relation between the type of gesture used during the explanation and performance on TOH2 in the *No-Switch group* simply because the smallest disk on TOH2 could be lifted using either one or two hands.

Beilock and Goldin-Meadow (2010) suggested that the one-handed gestures speakers produced during the explanation task helped to consolidate a representation of the smallest disk as "light." This representation was incompatible with the action that had to be performed on TOH2 in the *Switch group* but not in the *No-Switch group*. If gesturing is responsible for the decrement in performance in the *Switch group*, removing gesturing should eliminate the decrement—which is precisely what happened. In a second experiment that eliminated the explanation phase and thus eliminated gesturing entirely, the *Switch group* displayed no decrement in performance and, in fact, improved as much as the *No-Switch group* (Beilock & Goldin-Meadow, 2010). Thus, the switch in disks led to difficulties on TOH2 only when the adults gestured in between the two problem-solving attempts, and only when those gestures conveyed information that was incompatible with the speaker's next moves.

Note that disk weight is not a relevant factor in solving the Tower of Hanoi problem. Thus, when the speakers explained how they solved TOH1, they never talked about the weight of the disks or the number of hands they used to move the disks. However, it is difficult not to represent disk weight when gesturing—using a one-handed versus a two-handed gesture implicitly captures the weight of the disk, and this gesture choice had a clear effect on TOH2 performance. Moreover, the number of hands that the *Switch group* actually used when acting on the smallest disk in TOH1 did not predict performance on TOH2; only the number of one-handed gestures predicted performance. The findings suggest that gesture is adding action information to the speakers' mental representation of the task, rather than merely reflecting their previous actions. Gesturing about an action can thus solidify in mental representation the particular components of the action reflected in the gesture.

Conclusions and Future Directions

We know that the gestures speakers spontaneously produce along with their talk reflect their thoughts,

and that those thoughts are often not expressed in the talk itself. Moreover, evidence is mounting that gesture not only reflects thought but also plays a role in changing thought. The next frontier is to figure out *how* gesture influences thinking.

We have seen that gesture serves a range of functions for both listeners (communicative functions) and speakers (cognitive functions). One question for future research is how these functions work together. For example, gesturing reduces demands on the speaker's working memory, and it can also introduce new information into the speaker's mental representations. Are these two functions synergistic? Recall that producing gestures that convey different information from speech is particularly effective in lightening demands on the novice's working memory. Moreover, seeing gestures that convey different information from speech is highly effective in teaching the novice new information. The parallel hints at a potential relation between the two functions and warrants additional study.

Another important question is whether the processes that are responsible for the effect gesture has on learning are unique to gesture. Gesture may be special only in the sense that it makes efficient use of ordinary learning processes; for example, cues may be more distinctive when presented in two modalities than in one, and speech and gesture may simply be an effective way of presenting information multimodally. On the other hand, it is possible that traditional principles of learning and memory (e.g., distinctiveness, elaboration, cue validity, cue salience, etc.) will, in the end, not be adequate to account for the impact that gesture has on learning; in this event, it will be necessary to search for processes that are specific to gesture.

We have also seen that gesture is served by a range of mechanisms and has roots in speech, visuospatial thinking, and action. Again, the question is how these processes work together. If current theories are correct that speech has an action base (Barsalou, 1999; Glenberg & Kaschak, 2002; Richardson et al., 2003; Zwaan et al., 2002), gesture may be a natural reflection of this foundation. We can then ask whether action holds a privileged position not only in gesture production but also in gesture's effect on thinking. For example, do gestures that closely resemble action (e.g., simulating the movement of the hands as they lift the disks in the Tower of Hanoi task) have a more powerful effect on the mental representations of the speaker than gestures that incorporate some, but not all, aspects of the

action (e.g., tracing the trajectory of the disks as they are lifted, but including no information about how the hand was shaped as it moved the disk) or than gestures that are only abstractly related to action (i.e., metaphoric gestures)?

Another question for future research is whether gesture and action affect mental representations in the same way. Although gesturing is based in action, it is not a literal replay of the movements involved in action. Thus, it is conceivable that gesture could have a different impact on thought than action itself. Arguably, gesture should have less impact than action, precisely because gesture is "less" than action; that is, it is only a representation, not a literal recreation, of action. Alternatively, this "once-removed-from-action" aspect of gesture could have a more, not less, powerful impact on thought (see, for example, Goldin-Meadow & Beilock, 2010).

Finally, we can ask whether visuospatial thinking is privileged with respect to gesture. Gesture is an ideal medium for capturing spatial information, which leads to an important question about the mechanism that underlies gesture. Do domains that are inherently spatial (e.g., reasoning about the configuration of objects, as in organic chemistry) lend themselves to gesture more than nonspatial domains (e.g., reasoning about moral dilemmas)? Gesture's affinity with space also leads to questions about gesture's function. Does gesture affect cognitive processes more in spatial domains than in nonspatial domains? Is gesture effective in changing thinking because it can "spatialize" any domain (e.g., producing spatial gestures along with a description of a moral dilemma introduces spatial elements into the problem space and, in this way, allows spatial mechanisms to be brought to bear on the problem)?

The hope is that future work will allow us to build a model of exactly how speech and gesture emerge, both over ontogeny and in the moment during processing. In development, there is considerable evidence that early thoughts are often expressed in gesture prior to being expressed in speech, and that expressing those thoughts in gesture facilitates expressing them in speech. However, it is less clear whether gesture and speech have a similar relation during processing in the moment. Gestures often onset before the words they represent during production (McNeill, 1992) and, in fact, there is a precise relation in the timing of gesture to word—the more familiar the word, the smaller the gap between onset of gesture and onset of word (Morrel-Samuels & Krauss, 1992). But this timing relation need not reflect

the process by which thoughts are translated into gesture and speech. It is possible that some thoughts can be accessed by gesture before being accessed by speech (thoughts that are less amenable to speech and perhaps privileged in gesture). It is also possible that some thoughts are accessed by gesture only after they have been packaged into a spoken representation, although the results we have reviewed here demonstrating a direct link between gesture and thinking may make this second possibility less plausible.

In sum, the spontaneous gestures we produce when we talk are not mindless hand waving. They not only reflect our thoughts, but they also have the potential to change the thoughts of others (our listeners) and even to change our own thoughts (as speakers). Gesture thus offers a tool that allows both learners and researchers to make new discoveries about the mind.

Acknowledgments

The work described in this chapter was supported by NICHD Award PO1 HD406-05, NICHD Award R01 HD47450, NSF Award BCS-0925595, and NSF Award SBE 0541957 for the Spatial Intelligence Learning Center to SGM.

References

- Alibali, M. W., & Goldin-Meadow, S. (1993). Gesture-speech mismatch and mechanisms of learning: What the hands reveal about a child's state of mind. *Cognitive Psychology*, 25, 468–523.
- Alibali, M. W., Bassok, M., Solomon, K. O., Syc, S. E., & Goldin-Meadow, S. (1999). Illuminating mental representations through speech and gesture. *Psychological Science*, 10, 327–333.
- Alibali, M. W., & DiRussso, A. A. (1999). The function of gesture in learning to count: More than keeping track. *Cognitive Development*, 14, 37–56.
- Alibali, M. W., Heath, D. C., & Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language*, 44, 169–188.
- Alibali, M. W., Kita, S., & Young, A. J. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes*, 15, 593–613.
- Alverson, H. (1994). *Semantics and experience: Universal metaphors of time in English, Mandarin, Hindi, and Desotho*. Baltimore, MD: Johns Hopkins University Press.
- Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20, 723–767.
- Barbieri, F., Buonocore, A., Dalla Volta, R., & Gentilucci, M. (2009). How symbolic gestures and words interact with each other. *Brain and Language*, 110, 1–11.
- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577–660.
- Bavelas, J. B., Chovil, N., Lawrie, D. A., & Wade, A. (1992). Interactive gestures. *Discourse Processes*, 15, 469–489.
- Bavelas, J., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, 58, 495–520.
- Bavin, E. L., Prior, M., Reilly, S., Bretherton, L., Williams, J., Eadie, P., et al. (2008). The Early Language in Victoria study: Predicting vocabulary at age one and two years from gesture and object use. *Journal of Child Language*, 35, 687–701.
- Beattie, G., & Coughlan, J. (1999). An experimental investigation of the role of iconic gestures in lexical access using the tip-of-the-tongue phenomenon. *British Journal of Psychology*, 90, 35–56.
- Beattie, G., & Shovelton, H. (1999). Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech. *Journal of Language and Social Psychology*, 18, 438–462.
- Beattie, G., & Shovelton, H. (2000). Iconic hand gestures and the predictability of words in context in spontaneous speech. *British Journal of Psychology*, 91, 473–491.
- Beattie, G., & Shovelton, H. (2002). What properties of talk are associated with the generation of spontaneous iconic hand gestures? *British Journal of Social Psychology*, 41, 403–417.
- Beilock, S. L., & Goldin-Meadow, S. (2010). Gesture grounds thought in action. *Psychological Science*, 21, 1605–1610.
- Bernardis, P., & Gentilucci, M. (2006). Speech and gesture share the same communication system. *Neuropsychologia*, 44, 178–190.
- Bernardis, P., Salillas, E., & Caramelli, N. (2008). Behavioural and neurophysiological evidence of semantic interaction between iconic gestures and words. *Cognitive Neuropsychology*, 25, 1114–1128.
- Boncoddo, P., Dixon, J. A., & Kelley, E. (2010). The emergence of a novel representation from action: Evidence from preschoolers. *Developmental Science*, 13, 370–377.
- Booth, A. E., McGregor, K., & Rohlfing, K. (2008). Sociopragmatics and attention: Contributions to gesturally guided word learning in toddlers. *Language Learning and Development*, 4, 179–202.
- Boroditsky, L., & Ramscar, M. (2002). The roles of body and mind in abstract thought. *Psychological Science*, 13, 185–188.
- Broaders, S., Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2007). Making children gesture reveals implicit knowledge and leads to learning. *Journal of Experimental Psychology: General*, 136, 539–550.
- Cassell, J., McNeill, D., & McCullough, K. E. (1999). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics and Cognition*, 7, 1–34.
- Chawla, P., & Krauss, R. (1994). Gesture and speech in spontaneous and rehearsed narratives. *Journal of Experimental Social Psychology*, 30, 580–601.
- Chieffi, S., Secchi, C., & Gentilucci, M. (2009). Deictic word and gesture production: Their interaction. *Behavioral Brain Research*, 203, 200–206.
- Chu, M., & Kita, S. (2008). Spontaneous gestures during mental rotation tasks: Insights into the microdevelopment of the motor strategy. *Journal of Experimental Psychology: General*, 137, 706–723.
- Church, R. B., Ayman-Nolley, S., & Mahootian, S. (2004). The role of gesture in bilingual education: Does gesture enhance

- learning? *International Journal of Bilingual Education and Bilingualism*, 7, 303–319.
- Church, R. B., & Goldin-Meadow, S. (1986). The mismatch between gesture and speech as an index of transitional knowledge. *Cognition*, 23, 43–71.
- Clark, H. H. (1973). Space, time, semantics and the child. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (pp. 27–63). New York: Academic Press.
- Cohen, A. A. (1977). The communicative functions of hand illustrators. *Journal of Communication*, 27, 54–63.
- Cohen, A. A., & Harrison, R. P. (1973). Intentionality in the use of hand illustrators in face-to-face communication situations. *Journal of Personality and Social Psychology*, 28, 276–279.
- Cook, S. W., Jaeger, T. F., & Tanenhaus, M. K. (2009). Producing less preferred structures: More gestures, less fluency. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 62–67). Austin, TX: Cognitive Science Society.
- Cook, S. W., Mitchell, Z., Goldin-Meadow, S. (2008). Gesturing makes learning last. *Cognition*, 106, 1047–1058.
- Cook, S. W., & Tanenhaus, M. K. (2009). Embodied communication: Speakers' gestures affect listeners' actions. *Cognition*, 113, 98–104.
- Cook, S. W., & Goldin-Meadow, S. (2006). The role of gesture in learning: Do children use their hands to change their minds? *Journal of Cognition and Development*, 7, 211–232.
- Corballis, M. (1992). On the evolution of language and generativity. *Cognition*, 44, 197–226.
- Ehrlich, S. B., Levine, S. C., & Goldin-Meadow, S. (2006). The importance of gesture in children's spatial reasoning. *Developmental Psychology*, 42, 1259–1268.
- Ekman, P., & Friesen, W.V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1, 49–98.
- Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). *Emotion in the human face*. New York: Pergamon Press.
- Emmorey, K., & Casey, S. (2001). Gesture, thought and spatial language. *Gesture*, 1, 35–50.
- Feyereisen, P. (1997). The competition between gesture and speech production in dual-task paradigms. *Journal of Memory and Language*, 36, 13–33.
- Feyereisen, P., & Havard, I. (1999). Mental imagery and production of hand gestures while speaking in younger and older adults. *Journal of Nonverbal Behavior*, 23, 153–171.
- Fitch, W. T. (2000). The evolution of speech: A comparative review. *Trends in Cognitive Science*, 4, 258–267.
- Flöel, A., Ellger, T., Breitenstein, C., & Knecht, S. (2003). Language perception activates the hand motor cortex: Implications for motor theories of speech perception. *The European Journal of Neuroscience*, 1(8) 704–708
- Frick-Horbury, D., & Guttentag, R. E. (1998). The effects of restricting hand gesture production on lexical retrieval and free recall. *American Journal of Psychology*, 111, 44–61.
- Garber, P., & Goldin-Meadow, S. (2002). Gesture offers insight into problem-solving in adults and children. *Cognitive Science*, 26, 817–831.
- Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, 9, 558–565.
- Göksun, T., Hirsh-Pasek, K., & Golinkoff, R. M. (2010). How do preschoolers express cause in gesture and speech? *Cognitive Development*, 25, 56–68.
- Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Cambridge, MA: Harvard University Press.
- Goldin-Meadow, S., & Beilock, S. L. (2010). Action's influence on thought: The case of gesture. *Perspectives on Psychological Science*, 5, 664–674.
- Goldin-Meadow, S., & Butcher, C. (2003). Pointing toward two-word speech in young children. In S. Kita (Ed.), *Pointing: Where language, culture, and cognition meet* (pp. 85–107). Hillsdale, NJ: Erlbaum.
- Goldin-Meadow, S., Cook, S. W., & Mitchell, Z. A. (2009). Gesturing gives children new ideas about math. *Psychological Science*, 20, 267–272.
- Goldin-Meadow, S., Goodrich, W., Sauer, E., & Iverson, J. M. (2007). Young children use their hands to tell their mothers what to say. *Developmental Science*, 10, 778–785.
- Goldin-Meadow, S., Kim, S., & Singer, M. (1999). What the teacher's hands tell the student's mind about math. *Journal of Educational Psychology*, 91, 720–730.
- Goldin-Meadow, S., & Momeni Sandhofer, C. (1999). Gestures convey substantive information about a child's thoughts to ordinary listeners. *Developmental Science*, 2, 67–74.
- Goldin-Meadow, S., Nusbaum, H., Kelly, S., & Wagner, S. (2001). Explaining math: Gesturing lightens the load. *Psychological Science*, 12, 516–522.
- Goldin-Meadow S., & Singer, M. (2003). From children's hands to adults' ears: Gesture's role in the learning process. *Developmental Psychology*, 39, 509–520.
- Goodwyn, S., & Acredolo, L. (1993). Symbolic gesture versus word - is there a modality advantage for onset of symbol use. *Child Development*, 64, 688–701.
- Graham, J. A., & Argyle, M. (1975). A cross-cultural study of the communication of extra-verbal meaning by gestures. *International Journal of Psychology*, 10, 57–67.
- Graham, J. A., & Heywood, S. (1975). The effects of elimination of hand gestures and of verbal codability on speech performance. *European Journal of Social Psychology*, 2, 189–195.
- Graham, T. A. (1999). The role of gesture in children's learning to count. *Journal of Experimental Child Psychology*, 74, 333–355.
- Hadar, U., Burstein, A., Krauss, R., & Soroker, N. (1998). Ideational gestures and speech in brain-damaged subjects. *Language and Cognitive Processes*, 13, 59–76.
- Haviland, J. B. (1993). Anchoring, iconicity, and orientation in Guugu Yimithirr pointing gestures. *Journal of Linguistic Anthropology*, 3, 3–45.
- Hegarty, M., Mayer, S., Kriz, S., & Keehner, M. (2005). The role of gestures in mental animation. *Spatial Cognition and Computation*, 5, 333–356.
- Holden, C. (2004). The origin of speech. *Science*, 303, 1316–1319.
- Holle, H., & Gunter, T.C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience*, 19, 1175–1192.
- Holle, H., Obleser, J., Rueschemeyer, S., & Gunter, T. (2010). Integration of iconic gestures and speech in left superior temporal areas boosts speech comprehension under adverse listening conditions. *NeuroImage*, 49, 875–884.
- Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin and Review*, 15, 495–514.
- Hostetter, A. B., & Alibali, M. W. (2010). Language, gesture, action! A test of the gesture as simulated action framework. *Journal of Memory and Language*, 63, 245–257.

- Hostetter, A. B., Alibali, M. W., & Kita, S. (2007). I see it in my hands' eye: Representational gestures reflect conceptual demands. *Language and Cognitive Processes*, 22, 313–336.
- Hostetter, A. B., & Hopkins, W. D. (2002). The effect of thought structure on the production of lexical movements. *Brain and Language*, 82, 22–29.
- Iverson, J. M., & Goldin-Meadow, S. (1997). What's communication got to do with it? Gesture in children blind from birth. *Developmental Psychology*, 33, 453–467.
- Iverson, J. M., & Goldin-Meadow, S. (1998). Why people gesture when they speak. *Nature*, 396, 228.
- Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science*, 16, 367–371.
- Kelly, S. D. (2001). Broadening the units of analysis in communication: Speech and nonverbal behaviours in pragmatic comprehension. *Journal of Child Language*, 28, 325–349.
- Kelly, S. D., & Church, R. B. (1997). Can children detect conceptual information conveyed through other children's non-verbal behaviors? *Cognition and Instruction*, 15, 107–134.
- Kelly, S. D., & Church, R. B. (1998). A comparison between children's and adults' ability to detect conceptual information conveyed through representational gestures. *Child Development*, 69, 85–93.
- Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language*, 89, 253–260.
- Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, 21, 260–267.
- Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In M. Key (Ed.), *The relationship of verbal and nonverbal communication* (pp. 207–227). The Hague, Netherlands: Mouton.
- Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), *Language and gesture: Window into thought and action* (pp. 162–185). Cambridge, England: Cambridge University Press.
- Kita, S., & Davies, T. S. (2009). Competing conceptual representations trigger co-speech representational gestures. *Language and Cognitive Processes*, 24, 761–775.
- Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48, 16–32.
- Kita, S., Özyürek, A., Allen, S., Brown, A., Furman, R., & Ishizuka, T. (2007). Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. *Language and Cognitive Processes*, 22, 1212–1236.
- Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, 57, 396–414.
- Krauss, R. M. (1998). Why do we gesture when we speak? *Current Directions in Psychological Science*, 7, 54–60.
- Krauss, R. M., Dushay, R. A., Chen, Y., & Rauscher, F. (1995). The communicative value of conversational hand gestures. *Journal of Experimental Social Psychology*, 31, 533–552.
- Kutas, M., & Hillyard, S. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307, 161–163.
- Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. Chicago, IL: University of Chicago Press.
- Lakoff, G., & Nunéz, R. (2000). *Where mathematics comes from: How the embodied mind brings mathematics into being*. New York: Basic Books.
- Langton, S. R. H., O'Malley, C., & Bruce, V. (1996). Actions speak no louder than words: Symmetrical cross-modal interference effects in the processing of verbal and gestural information. *Journal of Experimental Psychology: Human Perception and Performance*, 22, 1357–1375.
- Larson, S. W., Ping, R. M., Zinchenko, E., Decatur, M., & Goldin-Meadow, S. (2010, August). *Adult gesture-speech mismatch predicts learning on a mental rotation task*. Poster presented at the Spatial Cognition International Conference, Mt. Hood, OR.
- Lavergne, J., & Kimura, D. (1987). Hand movement asymmetry during speech: No effect of speaking topic. *Neuropsychologia*, 25, 689–693.
- Levett, W. J. M., Richardson, G., & Laheij, W. (1985). Pointing and voicing in deictic expressions. *Journal of Memory and Language*, 24, 133–164.
- Luyster, R. J., Kedlec, M. B., Carter, A., & Tager-Flusberg, H. (2008). Language assessment and development in toddlers with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 38, 1426–1438.
- Masataka, N. (2001). Why early linguistic milestones are delayed in children with Williams syndrome: Late onset of hand banging as a possible rate-limiting constraint on the emergence of canonical babbling. *Developmental Science*, 4, 158–164.
- Mayberry, R. I., & Jacques, J. (2000). Gesture production during stuttered speech: Insights into the nature of gesture-speech integration. In D. McNeill (Ed.), *Language and gesture* (pp. 199–214). Cambridge, England: Cambridge University Press.
- McClave, E. (1998). Pitch and manual gestures. *Journal of Psycholinguistic Research*, 27, 69–89.
- McGregor, K. K., Rohlfing, K. J., Bean A., & Marschner, E. (2008). Gesture as a support for word learning: The case of under. *Journal of Child Language*, 36, 807–828.
- McNeill, D. (1992). *Hand and mind*. Chicago, IL: University of Chicago Press.
- McNeill, D., Levy, E., & Pedelty, L. (1990). Speech and gesture. In G. Hammond (Ed.), *Cerebral control of speech and limb movements*. (pp. 203–256). North Holland, Netherlands: Elsevier.
- McNeil, N. M., Alibali, M. W., & Evans, J. L. (2000). The role of gesture in children's comprehension of spoken language: Now they need it, now they don't. *Journal of Nonverbal Behavior*, 24, 131–150.
- Meister, I. G., Boroojerdi, B., Folts, H., Sparing, R., Huber, W & Töpper, R. (2003). Motor cortex hand area and speech: Implications for the development of language. *Neuropsychologia*, 41, 401–406.
- Melinger, A., & Kita, S. (2007). Conceptualisation load triggers gesture production. *Language and Cognitive Processes*, 22, 473–500.
- Morrel-Samuels, P., & Krauss, R. M. (1992). Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 615–622.
- Morsella, E., & Krauss, R. M. (2004). The role of gestures in spatial working memory and speech. *The American Journal of Psychology*, 117, 411–424.

- Morsella, E., & Krauss, R. M. (2005). Muscular activity in the arm during lexical retrieval: Implications for gesture–speech theories. *Journal of Psycholinguistic Research*, 34, 415–427.
- Morford, M., & Goldin-Meadow, S. (1992). Comprehension and production of gesture in combination with speech in one-word speakers. *Journal of Child Language*, 19, 559–580.
- Özçalışkan, S., & Goldin-Meadow, S. (2005). Gesture is at the cutting edge of early language development. *Cognition*, 96, B101–113.
- Özçalışkan, S., & Goldin-Meadow, S. (2009). When gesture–speech combinations do and do not index linguistic change. *Language and Cognitive Processes*, 24, 190–217.
- Özyürek, A., Kita, S., Allen, S., Furman, R., & Brown, A. (2005). How does linguistic framing of events influence co-speech gestures? Insights from crosslinguistic variations and similarities. *Gesture*, 5, 219–240.
- Özyürek, A., & Özçalışkan, S. (2000). How do children learn to conflate manner and path in their speech and gestures? Differences in English and Turkish. In E. V. Clark (Ed.), *The proceedings of the Thirtieth Child Language Research Forum* (pp. 77–85). Stanford, CA: CSLI Publications.
- Özyürek, A., Willems, R.M., Kita, S., & Hagoort, P. (2007). On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience*, 19, 605–616.
- Núñez, R., & Sweetser, E. (2006). With the future behind them: Convergent evidence from Aymara language and gesture in the crosslinguistic comparison of spatial construal's of time. *Cognitive Science*, 30, 401–450.
- Perry, M., Church, R. B., & Goldin-Meadow, S. (1988). Transitional knowledge in the acquisition of concepts. *Cognitive Development*, 3, 359–400.
- Perry, M., & Elder, A. D. (1997). Knowledge in transition: Adults' developing understanding of a principle of physical causality. *Cognitive Development*, 12, 131–157.
- Pine, K. J., Bird, H., & Kirk, E. (2007). The effects of prohibiting gestures on children's lexical retrieval ability. *Developmental Science*, 10, 747–754.
- Pine, K., Lufkin, N., Kirk, E., & Messer, D. (2007). A microgenetic analysis of the relationship between speech and gesture in children: Evidence for semantic and temporal asynchrony. *Language and Cognitive Processes*, 22, 234–246.
- Pine, K. J., Lufkin, N., & Messer, D. (2004). More gestures than answers: Children learning about balance. *Developmental Psychology*, 40, 1059–1067.
- Ping, R. M., & Goldin-Meadow, S. (2008). Hands in the air: Using ungrounded iconic gestures to teach children conservation of quantity. *Developmental Psychology*, 44, 1277–1287.
- Ping, R. M., & Goldin-Meadow, S. (2010). Gesturing saves cognitive resources when talking about nonpresent objects. *Cognitive Science*, 34, 602–619.
- Rauscher, F., Krauss, R., & Chen, Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science*, 7, 226–231.
- Richardson, D. C., Spivey, M. J., Barsalou, L. W., & McRae, K. (2003). Spatial representations activated during real-time comprehension of verbs. *Cognitive Science*, 27, 767–780.
- Rimé, B. (1982). The elimination of visible behavior from social interactions: Effects on verbal, nonverbal and interpersonal variables. *European Journal of Social Psychology*, 12, 113–129.
- Rimé, B., & Schiaratura, L. (1991). Gesture and speech. In R. S. Feldman & B. Rimé (Eds), *Fundamentals of nonverbal behavior* (pp. 239–281). New York: Cambridge University Press.
- Rime, B., Schiaratura, L., Hupert, M., & Ghysselinckx, A. (1984). Effects of relative immobilization on the speaker's nonverbal behavior and on the dialogue imagery level. *Motivation and Emotion*, 8, 311–325.
- Rogers, W. T. (1978). The contribution of kinesic illustrators toward the comprehension of verbal behavior within utterances. *Human Communication Research*, 5, 54–62.
- Rowe, M. L., & Goldin-Meadow, S. (2008). Early gesture selectively predicts later language learning. *Developmental Science*, 12, 182–187.
- Rowe, M. L., & Goldin-Meadow, S. (2009). Differences in early gesture explain SES disparities in child vocabulary size at school entry. *Science*, 323(5916), 951–953.
- Rowe, M. L., Özçalışkan, S., & Goldin-Meadow, S. (2008). Learning words by hand: Gesture's role in predicting vocabulary development. *First Language*, 28, 182–199.
- Sassenberg, U., & van der Meer, E. (2010). Do we really gesture more when it is more difficult? *Cognitive Science*, 34, 643–664.
- Sauer, E., Levine, S. C., & Goldin-Meadow, S. (2010). Early gesture predicts language delay in children with pre- or perinatal brain lesions. *Child Development*, 81, 528–539.
- Saxe, G. B., & Kaplan, R. (1981). Gesture in early counting: A developmental analysis. *Perceptual and Motor Skills*, 53, 851–854.
- Schwartz, D. L., & Black, T. (1999). Inferences through imagined actions: Knowing by simulated doing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 116–136.
- Schwartz, D. L., & Holton, D. L. (2000). Tool use and the effect of action on the imagination. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 1655–1665.
- Seyal, M., Mull, B., Bhullar, N., Ahmad, T., & Gage, B. (1999). Anticipation and execution of a simple reading task enhance corticospinal excitability. *Clinical Neurophysiology*, 110, 424–429.
- Singer, M., & Goldin-Meadow, S. (2005). Children learn when their teacher's gestures and speech differ. *Psychological Science*, 16, 85–89.
- Singer, M., Radinsky, J., & Goldman, S. R. (2008). The role of gesture in meaning construction. *Discourse Processes*, 45(4–5), 365–386.
- Silverman, L., Bennetto, L., Campana, E., & Tanenhaus, M. K. (2010). Speech-and-gesture integration in high functioning autism. *Cognition*, 115, 380–393.
- Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C., & Small, S. L. (2007). Speech-associated gestures, Broca's area, and the human mirror system. *Brain and Language*, 101, 260–277.
- Smith, V., Mirenda, P., & Zaidman-Zait, A. (2007). Predictors of expressive vocabulary growth in children with autism. *Journal of Speech Language and Hearing Research*, 50, 149–160.
- Sousa-Poza, J. F., Rohrberg, R., & Mercure, A. (1979). Effects of type of information (abstract-concrete) and field dependence on asymmetry of hand movements during speech. *Perceptual and Motor Skills*, 48, 1323–1330.
- Thal, D., & Tobias, S. (1992). Communicative gestures in children with delayed onset of oral expressive vocabulary. *Journal of Speech & Hearing Research*, 35, 1281–1289.

- Thal D., Tobias, S., & Morrison, D. (1991). Language and gesture in late talkers - a 1-year follow-up. *Journal of Speech and Hearing Research*, 34, 604–612.
- Thompson, L. A., & Massaro, D. W. (1994). Children's integration of speech and pointing gestures in comprehension. *Journal of Experimental Child Psychology*, 57, 327–354.
- Thompson, L. A., & Massaro, D. W. (1986). Evaluation and integration of speech and pointing gestures during referential understanding. *Journal of Experimental Child Psychology*, 42, 144–168.
- Tokimura, H., Tokimura, Y., Oliviero, A., Asakura, T., & Rothwell, J. C. (1996). Speech-induced changes in corticospinal excitability. *Annals of Neurology*, 40, 628–634.
- Trafton, J. G., Trickett, S. B., Stitzlein, C. A., Saner, L., Schunn, C. D., & Kirschbaum, S. S. (2006). The relationship between spatial transformations and iconic gestures. *Spatial Cognition and Computation*, 6, 1–29.
- Valenzeno, L., Alibali, M. W., & Klatzky, R. (2003). Teachers' gestures facilitate students' learning: A lesson in symmetry. *Contemporary Educational Psychology*, 28, 187–204.
- Wagner, S. M., Nusbaum, H., & Goldin-Meadow, S. (2004). Probing the mental representation of gesture: Is handwaving spatial? *Journal of Memory and Language*, 50, 395–407.
- Wesp, R., Hesse, J., Keutmann, D., & Wheaton, K. (2001). Gestures maintain spatial imagery. *American Journal of Psychology*, 114, 591–600.
- Wexler, M., Kosslyn, S. M., & Berthoz, A. (1998). Motor processes in mental rotation. *Cognition*, 68, 77–94.
- Willems, R. M., Özyürek, A., & Hagoort, P. (2007). When language meets action: The neural integration of gesture and speech. *Cerebral Cortex*, 17, 2322–2333.
- Wilson, N. L., & Gibbs, R. W., Jr. (2007). Real and imagined body movement primes metaphor comprehension. *Cognitive Science*, 31, 721–731.
- Wohlschlager, A., & Wohlschlager, A. (1998). Mental and manual rotation. *Journal of Experimental Psychology*, 24, 397–412.
- Wu, Y. C., & Coulson, S. (2005). Meaningful gestures: Electrophysiological indices of iconic gesture comprehension. *Psychophysiology*, 42, 654–667.
- Wu, Y. C., & Coulson, S. (2007). How iconic gestures enhance communication: An ERP study. *Brain and Language*, 101, 234–245.
- Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language comprehenders mentally represent the shape of objects. *Psychological Science*, 13, 168–171.

Impact of Aging on Thinking

Shannon McGillivray, Michael C. Friedman, and Alan D. Castel

Abstract

This chapter discusses the impact of aging on judgment and decision making, problem solving, reasoning, induction, memory, and metacognition, as well as the influence of expertise, training, and wisdom. In addition, the chapter presents theories of cognitive aging and addresses the ways in which changing goals (such as emotional goals) in old age can alter the processes and outcomes associated with cognitive operations. There is a wealth of research documenting age-related cognitive declines and impairments in areas such as decision making, reasoning, problem solving, category learning, and memory. However, in addition to addressing the potential difficulties older adults may experience when performing demanding cognitive operations, this chapter also examines certain situations and variables that have been shown to lessen or ameliorate age-related differences in performance. Lastly, the impact of training, expertise, and wisdom are discussed as they relate to successful cognitive aging.

Key Words: aging, cognitive aging, decision making, memory, metacognition, problem solving, reasoning, inductive learning, training, expertise, wisdom

As an increasingly large proportion of the population falls into the category of “senior citizen,” it is vital to understand and explore how aging impacts cognitive functioning. Even during normal, non-pathological aging (which is the exclusive focus of this chapter) there is a large amount of evidence that older adulthood is associated with a decline in certain cognitive abilities, some of which are summarized in Figure 33.1 (McCabe et al., 2010; see also Craik & Salthouse, 2008). As Figure 33.1 illustrates, there are substantial declines in working memory capacity, episodic memory, executive functioning, as well as the speed at which information is processed. However, as Figure 33.1 also shows, aging does not negatively impact all functions to the same degree, if at all; and there is growing evidence that potential age-related deficits are moderated by other important factors such as goals, motivation, and prior knowledge (e.g., Zacks & Hasher, 2006). The

current chapter will first discuss some of the major theories regarding age-related cognitive changes, as well as theories that address important changes during life-span development, more generally. This chapter will then review some classic as well as more recent findings within the areas of judgment and decision making, problem solving, reasoning, inductive learning, memory, and metacognition in older adults. In addition, the roles of emotion, expertise, training, and wisdom will be discussed as they relate to various aspects of cognition.

Cognitive Aging Theories

A number of theories have been proposed to explain why cognitive capabilities are so susceptible to the effects of aging. These theories focus on possible mechanisms driving age-related changes, and they highlight situations in which older adults are more or less likely to experience difficulties.

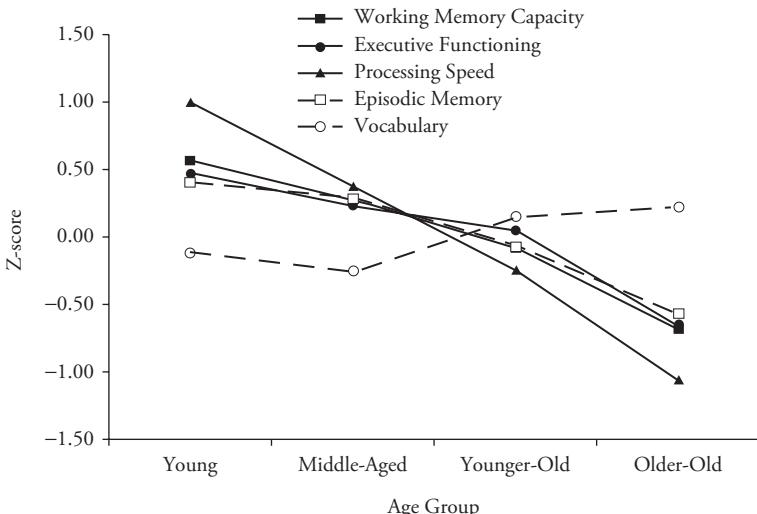


Fig. 33.1 Age-related differences in performance within various cognitive domains. The figure shows that aging is associated with declines in working memory capacity, executive functioning, processing speed, episodic memory, but an increase in vocabulary knowledge. (From McCabe, Roediger, McDaniel, Balota, & Hambrick, 2010. Copyright © 2010 by the American Psychological Association. Reproduced with permission.)

Although certainly not an exhaustive review of existing theories, this section discusses those that have received wide support within the cognitive aging literature: the general slowing theory, the reduced resources theory, the inhibition deficit theory, prefrontal theories, as well as the selective optimization with compensation theory and the socioemotional selectivity theory (both of which are more general theories of life-span development).

General Slowing Theory

The general slowing theory posits that a reduction in the speed with which cognitive processes operate occurs during aging (see Fig. 33.1, which shows a steep decline in processing speed), and this reduction in processing speed accounts for the majority of age-related variance on a variety of cognitive tasks (Henninger, Madden, & Huettel, 2010; Salthouse, 2000). For example, there is evidence that measures of speed share upward of 50%–75% of the age-related variance on numerous cognitive tasks (Salthouse, 1996). Salthouse (1996) suggests that there are two mechanisms responsible for the relationship between speed and cognition: limited time and simultaneity. Limited time plays an important role in that the time needed to perform later cognitive operations can become restricted when large portions of available time are taxed with earlier operations. Simultaneity refers to the idea that outputs of earlier cognitive processes may be

lost by the time that later processing is completed (as can occur when there are multiple demands on working memory), thus potentially creating situations in which relevant information is no longer available when it is actually needed.

Reduced Resources Theory

The reduced resource theory is similar to the general slowing theory, in that they both assert that a general change in specific cognitive abilities can account for large age-related changes in cognition. However, rather than positing a reduction in speed of processing, the reduced resources theory proposes that aging reduces the availability and/or the ability to successfully allocate attentional resources necessary for efficient performance on cognitive tasks (Craik & Byrd, 1982). For example, when older adults are placed under divided attention (which reduces the amount of attention available for other tasks), there is a larger detrimental impact on performance compared with younger adults also under divided attention (Anderson, Craik, & Naveh-Benjamin, 1998; Park, Smith, Dudley, & Lafronza, 1989). The reduction in available attentional resources can make it difficult for older adults to engage in more cognitively demanding operations, such as elaborative encoding during memory operations, which is considered necessary for effective consolidation and retrieval of to-be-remembered information (Craik & Salthouse, 2008).

Inhibition Deficit Theory

While there is evidence for age-related general cognitive slowing and a reduction in resources such as attention (which limit the amount of information one can process), other theories have proposed that older adults' troubles stem from the processing of too much (irrelevant) information. Hasher and colleagues (Darowski, Helder, Zacks, Hasher, & Hambrick, 2008; Hasher & Zacks, 1988; Lustig, Hasher, & Zacks, 2007) have suggested that older adults may suffer disproportionately from deficits in inhibitory processes (inhibition deficit theory), and this, in turn, can lead to poorer performance on cognitive tasks. An efficient system requires control and inhibition of irrelevant information in order to function properly, and thus it requires working memory and attention. Older adults in particular may have difficulty suppressing inappropriate or irrelevant responses, controlling the focus of attention, and keeping irrelevant information out of working memory and the focus of attention. As Figure 33.2 depicts, inefficient inhibition, therefore, can lead to information unrelated to the "goal path" entering working memory, resulting in a disruption of task operations. These non-goal path thoughts can involve irrelevant environmental details, personal memories and concerns, and interpretations that are inconsistent with current goals. Furthermore, decreased inhibitory functions can reduce the ability to switch attention from one target to another, and it can lead to misinterpretation of information, inappropriate responses, and also forgetting.

Prefrontal Theories

From a more neurological perspective, there is evidence that the prefrontal regions of the brain, which are responsible for many higher order cognitive operations (see Morrison & Knowlton, Chapter 6), are particularly susceptible to age-associated atrophic

changes (Cabeza, 2001; Raz et al., 1997). Such specific, age-related changes in the prefrontal cortex likely contribute to cognitive decline in older adults (West, 1996). In particular, performance on tasks reliant on dorsolateral prefrontal function (e.g., executive functioning and working memory) seem to be the most negatively affected during the normal aging process, whereas tasks associated with ventromedial prefrontal areas (e.g., social behavior regulation and emotion) are less affected (MacPherson, Phillips, & Della Sala, 2002). Furthermore, it has been suggested that it is the compromised integrity of not only the dorsolateral prefrontal regions but the dopamine projections to the prefrontal cortex that contribute to age-related cognitive declines (Braver & Barch, 2002). For example, there is evidence that dopamine and dorsolateral dysfunction contribute largely to older adults' deficits on tasks in which cognitive control is necessary, such as efficient inhibitory control, working memory, and attention (Braver & Barch, 2002).

Selective Optimization With Compensation Theory

In addition to theories that focus on the mechanisms driving declines in cognitive function during aging, there are also theories that explore the contributing factors to successful cognitive aging. Selective optimization with compensation (SOC; Baltes & Baltes, 1990) asserts that successful aging is related to a focused and goal-directed investment of limited resources into areas that yield optimal returns. Thus, older adults can selectively choose certain options in order to maximize performance based on their goals, compensating for impairments by optimizing performance in these specific, goal-related domains (see also Riediger, Li, & Lindenberger, 2006, for the adaptive nature of SOC). As the SOC theory suggests, older adults are able to successfully allocate limited resources when appropriate motivation (such

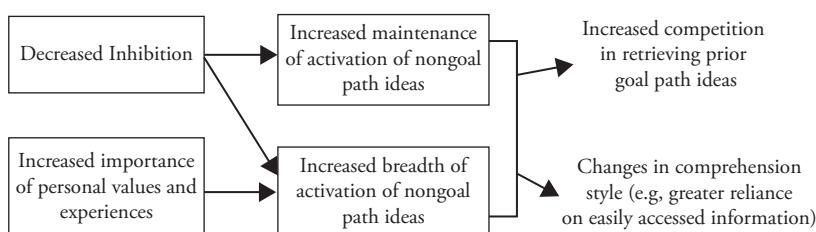


Fig. 33.2 Theoretical framework depicting the consequences of reduced inhibitory control as well as changing goals. (Adapted from *Psychology of Learning and Motivation*, Vol. 22, Hasher & Zacks, "Working Memory, Comprehension, and Aging: A Review and a New View," p. 213, Copyright 1988, with permission from Elsevier.)

as personal relevance and accountability) is present, enhancing performance (Germain & Hess, 2007; Hess, Rosenberg, & Waters, 2001). Furthermore, Heckhausen (1999; Heckhausen & Schulz, 1995) suggests that individuals have to take on the regulation of losses in aging-related resource in order to function efficiently, and if successful, such regulation can aid efficient cognitive function.

Socioemotional Selectivity Theory

Lastly, although not a theory of cognitive aging per se, the socioemotional selectivity theory (SST; Carstensen, 1992; 1995) highlights the importance of changing goals and motivations during aging. The SST asserts that people have some sort of awareness of the time left in life, and when time is seen as open ended (as it may be for young, healthy adults), goals and motivations are focused on acquiring information, experiencing novelty, and expanding one's knowledge. When time is seen as more limited (as may be the case for older adults), motivation and goals focus more on monitoring the environment in order to optimize emotional meaningfulness and emotional functioning. Depictions of the trajectories of these changing motivations are displayed in Figure 33.3, which show that in middle-to-older adulthood social motives shift from being more knowledge driven to more emotionally driven. Evidence supporting this theory has shown that older adults are better than younger adults at regulating emotions (Carstensen, Pasupathi, Mayr, & Nesselroade, 2000), prefer to spend time with more emotionally meaningful (compared to novel) social partners (Fredrickson & Carstensen, 1990), and are more likely to remember information emphasizing emotional relative to novelty-seeking information (Fung, Carstensen, & Lutz, 1999). Thus, while this framework is not a specific theory of cognitive aging, it has implications for the approach that older adults may take toward decision making, problem solving,

remembering information, and achieving emotional goals.

Summary

The purpose of this brief and selective review of theories regarding cognitive aging was to bring to the fore some of the possible mechanisms driving age-related changes in cognition. As was discussed, older adults may experience difficulties on cognitive tasks due to decreases in the speed (and thus efficiency) with which cognitive processes operate, decreases in the availability of attentional resources and/or in the ability to effectively allocate attention, and decreased ability to successfully inhibit irrelevant and intrusive competing information. Furthermore, age-related cerebral atrophy occurs at disproportionately higher rates within regions of the frontal lobe, an area that is largely responsible for many of the higher order cognitive operations. At the same time, life-span theories of aging suggest that older adults may approach tasks and situations in a qualitatively different manner than younger adults (e.g., older adults may have different goals) and, at times, can selectively allocate resources in order to compensate for deficiencies in cognitive abilities. It is important to consider these theories, and other potential frameworks, as we now review and discuss the effects of age within specific areas of cognition.

Judgment and Decision Making

As individuals age they are faced with a number of life changes and often need to make important decisions involving aspects such as medical care and retirement, in addition to the everyday decisions faced by most individuals. Thus, an understanding of judgment and decision-making abilities in older adults is of paramount importance. Research has suggested that the decision-making ability of older adults in everyday life may be compromised

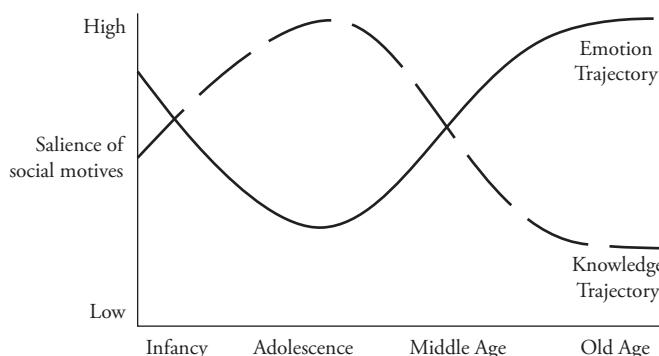


Fig. 33.3 Idealized depiction of changes in two social motives across the life span as predicted by the socioemotional selectivity theory. (Adapted from Carstensen, Gross, & Fung. Copyright 1997 by Springer Publishing Company Inc. Reproduced with permission of Springer Publishing Company, Inc.)

relative to younger and middle-aged adults (Peters, Finucane, MacGregor, & Slovic, 2000; Thornton & Dumke, 2005). In particular, older adults may make more comprehension errors and display less consistency in their preferences (Finucane et al., 2002), and they may exhibit poorer decision-making abilities when the task requires more cognitively demanding strategies (Mata, von Helversen, & Rieskamp, 2010). Furthermore, a decline in more controlled, deliberative processing and an increased reliance on more automatic, heuristic processing may lead to less effective or poorer decisions among older adults (dual-process model; see Peters, Hess, Västfjäll, & Auman, 2007; also Evans, Chapter 8). In this section we will highlight not only the instances in which age-related deficits are observed but also situations in which deficits are not present. We will also consider the impact of goals and motivations on the aging decision maker.

In the real world, individuals are often charged with making decisions in the face of gains, losses, risks, and uncertainty; the ability to decide advantageously in these situations is of great importance. Laboratory-based studies of decision making in the presence of the aforementioned factors have frequently utilized the Iowa Gambling Task (IGT), originally developed by Bechara and colleagues (Bechara, Damasio, Damasio, & Anderson, 1994), which presents individuals with the opportunity to either gain or lose large monetary amounts. In the IGT, there are four decks of cards, and each card has either a positive (gain) or negative (loss) monetary

amount on the reverse side. Subjects are allowed to select cards freely, one at a time, from any deck, with the task ending once they have selected 100 cards. There are always two “good” decks, which consist of smaller immediate gains and lower overall losses, and two “bad” decks, which contain very large gains but also larger long-term losses. To be successful on this task, one must learn to choose predominately from the “good” decks and avoid the “bad” decks.

Several studies suggest that, on average, older adults are more likely to make disadvantageous decisions on the IGT compared with younger adults (Denburg et al., 2007; Denburg, Tranel, & Bechara, 2005; Fein, McGillivray, & Finn, 2007). Figure 33.4 displays the average number of cards selected from the good decks (i.e., decks C and D) minus the number chosen from the bad decks (i.e., decks A and B) for both younger and older adults across blocks of 20 cards. The pattern clearly shows less advantageous decision making by older adults. Furthermore, even studies that have reported few age differences on this task have found overall flatter learning curves among older individuals (i.e., they take longer to adopt a successful strategy; see Wood, Busemeyer, Koling, Cox, & Davis, 2005). It may be the case that older adults are more sensitive to gains and less sensitive to losses than younger adults (Friedman, Castel, McGillivray, & Flores, 2010; Samanez-Larkin et al., 2007), which could partially account for their lower overall performance on the IGT. That is, individuals who place more emphasis on the larger gains, and deemphasize the larger

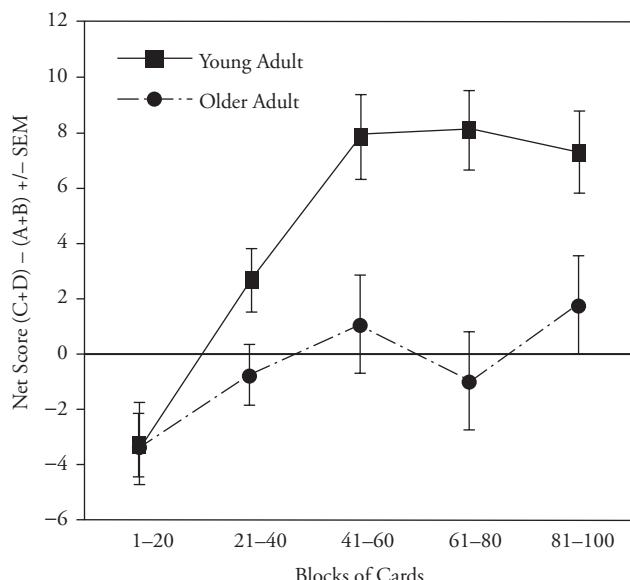


Fig. 33.4 The average number of cards selected from the good decks (C and D) minus the bad decks (A and B) across blocks of 20 cards for both younger and older adults. (From Denburg et al., 2007, reprinted with permission from John Wiley and Sons.)

losses, may continue to select cards from the “bad” decks. Alternatively, it has been suggested that these age-related differences on the IGT may in fact be a result of, or mediated by, declines in other cognitive abilities such as processing speed and explicit memory function (Henninger et al., 2010). What is particularly concerning, however, is that older individuals who perform poorly on the IGT are also more likely to fall prey to deceptive and fraudulent advertising (Denburg et al., 2007). What makes this alarming is the fact that older adults are often preferentially targeted by fraudulent schemes (American Association of Retired Persons, 1996), and thus some older adults may be at an increased risk of becoming victims of such crimes.

Despite evidence suggesting age-related deficits in decision making, a number of studies have found that older adults’ decision-making ability is not always compromised (Kim & Hasher, 2005; Kovalchik, Camerer, Grether, Plott, & Allman, 2005). Older adults, if given the opportunity, can become more adaptive and adopt different compensatory strategies (Mata, 2007), consistent with the previously discussed selective optimization with compensation theory. For instance, although older adults often review less information and take longer to process it, older adults are capable of adopting more complex, less heuristic-based decision-making strategies when the environment requires the use of such strategies (Mata, Schooler, & Rieskamp, 2007; Pachur, Mata, & Schooler, 2009) or if properly motivated (Kim, Goldstein, Hasher, & Zacks, 2005).

To illustrate this point, consider the framing effect, a phenomenon in which decisions and choices are altered by the way in which the options are presented. The most commonly known example is the Asian disease problem originally examined by Tversky and Kahneman (1981). When presented with this problem, individuals are more likely to demonstrate risk seeking in their choice when the options are framed as losses (400 out of 600 people will die), and risk aversion when the options are framed in terms of gains (200 out of 600 people will be saved; Tversky & Kahneman, 1981). If individuals rely on more automatic, heuristic processing when presented with these types of problems, they are more likely to fall victim to the framing effect. Thus, it is not surprising that older adults, who may rely on more heuristic-based processing styles due to limited resources, show larger framing effects than younger adults (Kim et al., 2005). However, when asked to provide justification for their choices

(i.e., have appropriate motivation), older adults (and younger adults) adopted a more systematic processing and no longer showed susceptibility to the language in which options were framed (Kim et al., 2005). It is also important to note that regardless of whether older adults review less information and rely on more heuristic-based processing when making decisions, reviewing less information does not necessarily lead to poorer quality of decisions (Mata & Nunes, 2010). In addition, recent work has shown that older adults actually prefer fewer choices when making decisions, and that their performance is related to numerical processing ability, general slowing, and working memory capacity (e.g., Peters et al., 2007; Reed, Mikels, & Simon, 2008; Tanius, Wood, Hanoch, & Rice, 2009).

While motivation certainly plays a role in enhancing the decision quality of older adults, goals (emotional and social goals in particular) influence decision making and choice evaluation in key ways as individuals age. Studies examining decision making within emotional or social realms have often found no age-related difference (MacPherson et al., 2002). As proposed by the socioemotional selectivity theory (SST), emotional regulation and enhancing emotional well-being may be important goals for older adults. If this is the case, then older adults may process and remember information surrounding decisions differently than do younger adults. For example, on a task assessing decisions for health care plans and doctors, it was found that older adults reviewed a greater proportion of positive compared to negative material than did younger adults, and they remembered the doctors and health plans they had chosen more positively (Lockenhoff & Carstensen, 2007). Additional studies have also found that older adults often remember their decision choices as having more positive features (Mather & Johnson, 2003) and are more satisfied with their decisions compared with younger adults (Kim, Healey, Goldstein, Hasher, & Wiprzycka, 2008).

The SST has further proposed that older adults may see time as more limited than do younger adults and thus may see their time as more “valuable.” Given this hypothesis, older adults may be less susceptible to sunk cost effects, a common decision-making bias in which people continue to invest either time or money when prior investments have been made, despite limited prospects for positive returns (see LeBoeuf & Shafir, Chapter 16). In one study, older and younger adults were presented with scenarios in which, for example, they

had paid a certain amount of money to see a movie. Participants were told to imagine that shortly into the movie, they realized that it was not very good and were not enjoying it, and were given options to stop watching or continue watching for various lengths of time (Strough, Mehta, McFall, & Schuller, 2008). Strough and colleagues found that older adults were much less likely to demonstrate the sunk cost bias (i.e., continue watching the movie) than were younger adults. Thus, at least under some circumstances older adults may make more rational, “normatively correct” decisions than do younger individuals, possibly reflecting greater consideration of whether their previous investments of either time or money are currently yielding optimal positive returns.

Summary

Research has documented substantial deficits in judgment and decision making in older adults. For example, older adults make less advantageous decisions, rely more on heuristics when forming judgments and making decisions, and often review less information compared with younger adults. However, when appropriate motivation is present, age-related differences can be reduced and older adults process information in a more deliberate and appropriate manner. Changes in emotional goals may lead older adults to focus on more positive compared to negative aspects related to the decision-making process. Lastly, changes in time horizons (i.e., awareness of limited time left in life) can, in specific instances, result in more appropriate decisions by older adults.

Problem Solving, Reasoning, and Induction

The ability to solve problems, reason logically, classify objects into appropriate categories, and to sensibly come to novel conclusions based on known facts or rules are all critical abilities required to successfully manage one’s way through life. Research presented in the following sections will address age-related declines within these specific domains (e.g., Salthouse, 2005; Thornton & Dumke, 2005), along with the impact that goals, strategies, and prior knowledge have on older adults’ performance. In addition, strategies and instances that have a positive effect on older adults’ abilities will be examined.

Problem Solving

There is evidence that the ability to effectively solve problems (see Bassok & Novick, Chapter 21)

is reduced in later adulthood. For example, a meta-analysis of 28 separate studies concluded that problem solving is not spared from typical age-related declines (Thornton & Dumke, 2005). Older adults’ performance is lower than other age groups on both traditional, laboratory-based problems (Denney & Palmer, 1981), practical problems (Denney, Pearce, & Palmer, 1982), and even on problems specifically designed to give older adults an experience advantage (e.g., what an elderly woman should do if she needs to go somewhere at night, but she cannot see well enough to drive at night and it’s too far to walk; Denney & Pearce, 1989). Older adults may also review less information and generate fewer strategies during problem solving (Berg, Meegan, & Klaczynski, 1999). Furthermore, these deficiencies in everyday problem-solving tasks among older adults have been associated with measures of executive functioning, memory, verbal ability, and speed of processing (Burton, Strauss, Hultsch, & Hunter, 2006).

Despite difficulties with problem solving later in life, it may be that practical, realistic problems differ from the types of tasks used in typical laboratory settings. Although aging has been shown to negatively impact everyday problem-solving abilities (Thornton & Dumke, 2005), these age-related differences are much smaller than those observed on traditional problem-solving tasks (Denney & Palmer, 1981), suggesting that life experience and prior knowledge can moderate age-related declines. For example, Crawford and Channon (2002) gave younger and older adults a range of everyday situations that presented problems for which they needed to generate potential solutions for (e.g., resolving an issue with a neighbor’s barking dog). They found that while older adults generated fewer solutions compared with younger adults, these solutions were of a higher quality, which could be attributed to their greater life experience in dealing with these types of everyday problems (similar to findings regarding the role of expertise late in life; see Charness, 1981a, 1981b). That is, Crawford and Channon suggest that older adults may have a more well-defined knowledge base from which to draw possible solutions and make more efficient use of such knowledge compared to younger adults.

While typical problem-solving experiments are tested in a laboratory or controlled setting, problem solving in “real-life” occurs in a more complex and social environment. Many older adults talk about their problems with friends and family members, which has led to several studies within the last

decade exploring problem solving with older adults in the context of collaboration (Cheng & Strough, 2004; Kimbler & Margrett, 2009; Strough, Cheng, & Swenson, 2002; Strough, Hicks-Patrick, Swenson, Cheng, & Barnes, 2003; Strough, McFall, Flinn, & Schuller, 2008). For example, Cheng and Strough (2004) had either individual or collaborative same-sex pairs of younger and older adults plan a cross-country trip to go to a wedding. Although younger adults took less time and performed better at planning the trip overall, collaborating in a pair was advantageous for both age groups to the same extent, illustrating another strategy older individuals may employ to maintain everyday functioning.

As some of the previously mentioned research suggests, older adults may approach interpersonal problems in a qualitatively different manner than younger adults. It has been shown that older adults may use more effective problem-solving strategies than younger adults when faced with problems that are interpersonal in nature (e.g., conflicts with friends or family; Blanchard-Fields, Mienaltowski, & Seay, 2007). Furthermore, a number of studies have highlighted the fact that older adults are more likely to use (Blanchard-Fields, Chen, & Norris, 1997; Blanchard-Fields, Jahnke, & Camp, 1995) and prefer (Watson & Blanchard-Fields, 1998) emotionally focused problem-solving strategies compared with younger adults, particularly within interpersonal contexts, although both groups tend to use problem-focused strategies more often overall (Blanchard-Fields et al., 1995). One reason why older adults may use emotionally focused strategies more often than younger individuals could stem from differences in goals between the two age groups. It has previously been suggested that maintaining emotional well-being is an important goal for older individuals (e.g., Carstensen, 1992). It has been established that prioritization of emotional regulation has a sizable influence on the types of problem-solving strategies that are likely to be utilized by older adults (Coats & Blanchard-Fields, 2008; Hoppmann, Coats, & Blanchard-Fields, 2008). That is, older adults are more likely to endorse more passive emotional regulation strategies when solving interpersonal problems (Blanchard-Fields et al., 1997; Blanchard-Fields et al., 2007; Blanchard-Fields, Stein, & Watson, 2004; Coats & Blanchard-Fields, 2008), possibly due to their desire to maintain emotional stability and balance, particularly within their interpersonal relationships.

Lastly, it is important to note that areas such as everyday problem solving are a multidimensional construct, often with little relation between the different measures used (Allaire & Marsiske, 2002; Marsiske & Willis, 1995). Furthermore, performance is also modulated by such factors as education (Thornton & Dumke, 2005) and health (Diehl, Willis, & Schaie, 1995). In addition to the factors mentioned earlier, older adults, at times, do perform better when faced with a situation relevant to their own age group, analogous to an “own-age” bias (Artistico, Orom, Cervone, Krauss, & Houston, 2010). Other factors such as positive feedback (Soederberg-Miller & West, 2010), experience, and strategic flexibility have also been shown to improve older adults’ problem-solving and decision-making abilities (Hicks-Patrick & Strough, 2004).

Reasoning

Similar to findings observed in problem solving, the capability to effectively reason is negatively impacted during aging. Difficulties in reasoning ability are apparent by looking at older adults’ performance on the Raven’s Progressive Matrices Task, which shows a clear decline across the adult life span (Salthouse, 1993, 1994; Salthouse & Skovronek, 1992). The Raven’s task requires participants to identify an appropriate option to fill in a missing cell on a matrix grid that becomes progressively more difficult across trials. Figure 33.5 contains examples of the Raven’s Matrices (Fig. 33.5a), displaying problems of varying difficulty (i.e., the number of relations within the problem). A summary of older and younger adults’ performance on this task (Fig. 33.5b) reveals significant age-related differences at all levels of difficulty.

In addition to Raven’s Matrices, age-related performance deficits have also been observed on other reasoning tasks. For example, older adults performed worse than younger adults on a propositional reasoning task that presented individuals with a series of premise pairs (e.g., A > B, B > C, C > D), and then required them to draw inferences about a new pair (e.g., A ? C) (Ryan, Moses, & Villate, 2008). On tasks assessing analogical reasoning, older adults perform less accurately compared with younger and middle-aged adults (Viskontas, Morrison, Holyoak, Hummel, & Knowlton, 2004). Viskontas and colleagues found that this age-related deficit in analogical reasoning was present even at low levels of relational complexity, and it became more pronounced when problems contained an increasing

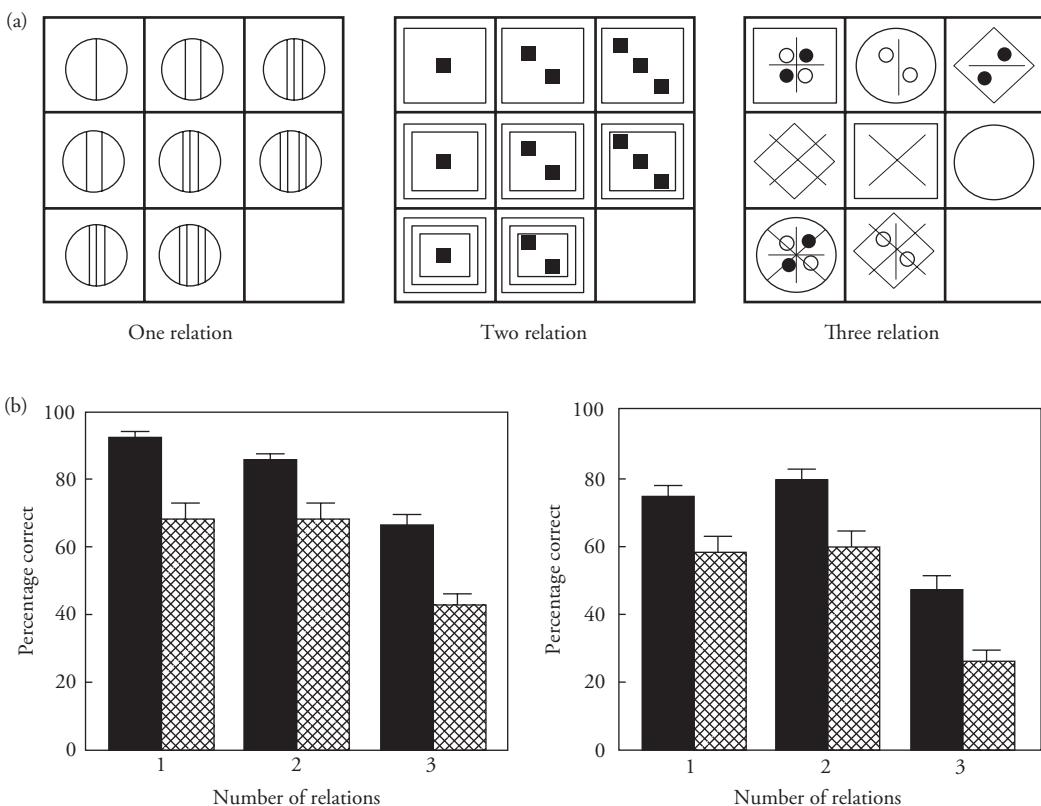


Fig. 33.5 a. A series of example patterns from Raven's Progressive Matrices Task with one, two, or three relations. Participants were given each matrix one at a time and asked to complete the missing cell with the appropriate pattern. Note that as the number of relations within a matrix increases, the task becomes more difficult to solve. b. Results for Raven's Progressive Matrices in younger adults (black bars) and older adults (gray bars). The results show significant age-related differences for both simultaneous (*left*) as well as sequential (*right*) conditions. Error bars indicate standard error of the mean. (Figures 33.5a and 5b taken from Salthouse, 1993 and reproduced with permission the British Journal of Psychology © The British Psychological Society.)

number of irrelevant traits that favored incorrect responses. Age-related declines in reasoning ability have been attributed to multiple sources, including general slowing (Salthouse, 2000), neurological changes to the prefrontal cortex (Krawczyk et al., 2008), differences in relational organization (Ryan et al., 2008), inhibitory decrements (Viskontas et al., 2004), and deficits in working memory (Kyllonen & Christal, 1990; Viskontas, Holyoak, & Knowlton, 2005).

Despite fundamental age-related differences in reasoning performance, the same mechanism may drive reasoning in both younger and older adults. For example, although older adults performed worse than younger adults on the Raven's Progressive Matrices Task, the types of errors both age groups committed (e.g., failure to identify all of the relevant variables needed to determine the correct solution, misunderstanding that some elements of the problem are not relevant to the solution) were

similar to one another, suggesting similar underlying mechanisms (Babcock, 2002; Salthouse, 1993). Furthermore, it has been found that these age-related deficits in reasoning can be reduced when using existing relational information within semantic memory (e.g., prior knowledge) as an analog for new learning (Ostreicher, Moses, Rosenbaum, & Ryan, 2010). There may even be specific situations in which older adults' lifetime of acquired wisdom and experiences result in superior reasoning abilities compared with younger adults (Grossmann et al., 2010; we return to this issue in the section on "Wisdom and Successful Aging").

Induction

Induction, or category learning (the ability to successfully place novel stimuli into one or more appropriate groups; see Rips, Smith, & Medin Chapter 11), has also been found to be susceptible to age-related declines (e.g., Filoteo & Maddox,

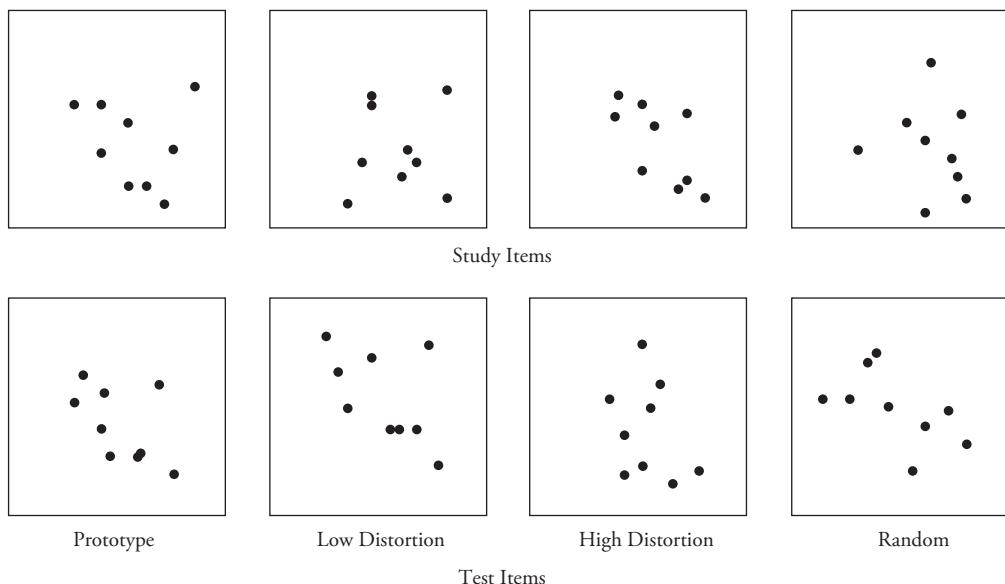


Fig. 33.6 Examples of dot patterns that are presented during study (*top*) and test (*bottom*). The study items are all distortions of the prototype dot pattern. The test items include the training prototype, high and low distortions of the prototype, and random dot patterns.

2004; Racine, Barch, Braver, & Noelle, 2006). One classic task used to assess category learning is the dot classification task (Posner & Keele, 1968). In this task participants are shown a series of dot patterns that are distortions of a predetermined prototypical pattern, examples of which are displayed at the top of Figure 33.6. Participants view a series of these patterns, and then during test are given the prototype pattern, high and low distortions of the prototype, as well as random dot patterns and are asked to categorize them (example test stimuli are displayed in the bottom portion of Fig. 33.6). Compared with younger adults, older adults have been shown to be less successful at correctly categorizing the test stimuli (Davis, Klebe, Bever, & Spring, 1998), and they retain less information about exemplars presented in the set (Hess & Slaughter, 1986). This increased difficulty in inductive learning tasks has been linked to general cognitive slowing, which can make it more difficult for older adults to successfully integrate information (Henninger et al., 2010; Mutter & Plumlee, 2009; Salthouse, 2000).

Similar to the processes involved in induction, older adults have been found to have difficulty in set-shifting (Ridderinkhof, Span, & van der Molen, 2002) and to exhibit more errors (Boone, Ghaffarian, Lesser, & Hill-Gutierrez, 1993; Rhodes, 2004) relative to younger adults on the Wisconsin Card Sorting Task. Although older adults are capable of

learning rules, the inability to appropriately switch or think “flexibly” may explain their reduced performance on induction tasks (see also Friedman & Castel, 2010; Koutstaal, 2006).

Despite increased difficulties with induction, older adults may benefit from specific learning parameters during the induction process. For example, Kornell, Castel, Eich, and Bjork (2010) demonstrated that although older adults did worse than younger adults on an assigned induction task—correctly identifying novel paintings from artists they had previously studied—both age groups benefitted from the same schedule of learning. Specifically, older and younger adults’ performance increased when exemplars from an artist were presented spaced further apart rather than massed together during the learning phase. Thus, spacing benefitted induction across both age groups, suggesting that mechanisms supporting inductive learning stay constant during aging (see also Jamieson & Rogers, 2000). Furthermore, Blieszner, Willis, and Baltes (1981) demonstrated that the ability to modify inductive learning and reasoning performance through interventions and training remains intact across the adult life span.

Summary

The areas of problem solving, reasoning, and inductive learning are subject to sizable age-related declines. It is well documented that older adults

demonstrate declines on both traditional, and to a lesser extent, everyday problem-solving tasks. In addition, much of the research conducted on older adults' reasoning capabilities reveal large age-associated decrements, even on tasks with relatively low levels of relational complexity. Although there is less research on inductive learning and aging, the existing literature supports the conclusion that older adults' capacity to learn categories is also compromised, and that they have difficulty learning categories that require rapid updating and the incorporation of new changing rules (such as on the Wisconsin Card Sorting task). However, there appear to be some contexts in which older adults are capable of performing just as well, if not better, than younger adults despite age-related cognitive declines. For example, older adults may generate fewer solutions to problems, but the quality of the solutions generated can, at times, be on par with younger adults. There is also evidence that older adults approach interpersonal problems in a qualitatively different way than younger adults, and their ability within this area of problem solving remains relatively intact. Furthermore, reasoning and induction tasks that allow older adults to utilize prior knowledge and experience also tend to show fewer age-related deficits.

Memory and Metacognition

Memory

Our memory is a vital component of who we are as individuals, and it allows us to efficiently interact with and understand the world. Not only do our memories contain information about our past experiences and what we know, but they influence our current and future actions. While there are many physical and psychological changes that accompany the aging process, one of the most oft-voiced concerns among many older adults is the decline in memory functioning. In fact, 50%–80% of older adults report subjective memory complaints (Levy-Cushman & Abeles, 1998). Older adults' subjective experience of memory difficulties has proven to be a well-founded concern, with many decades of research demonstrating that memory functioning declines with advancing age (e.g., Craik & Salthouse, 2008; Kausler, 1994). It is important to note, however, that there are numerous "types" of memory (e.g., episodic, semantic, working, procedural) and, as was shown earlier in Figure 33.1, aging may disproportionately impact these types of memory, with some, but not all, tasks associated

with age-related deficits (Craik & Salthouse, 2008; Kausler, 1994; Zacks & Hasher, 2006). Implicit or nondeclarative types of memory such as priming, skill learning, and classical conditioning, which rely more on automatic processes, generally show little to no age-related declines (e.g., Fleischman, Wilson, Gabrieli, Bienias, & Bennett, 2004; Laver, 2009; Light & Singh, 1987; Nilsson, 2003). Furthermore, semantic memory (i.e., memory for facts, world knowledge) is well preserved across the life span and in some instances, such as vocabulary knowledge (as is shown in Fig. 33.1) may even increase slightly (Lavoie & Cobia, 2007; Verhaeghen, 2003). Unlike implicit and semantic memory, however, large age-related declines are often observed in assessments of episodic memory (i.e., memory for past events) and working memory (i.e., short-term storage and manipulation of information; Verhaeghen & Salthouse, 1997).

It has been suggested that older adults' decline in explicit memory abilities can be attributed to declines in processing speed (Salthouse, 1996), attentional deficits (Craik & Byrd, 1982), and inefficient inhibitory mechanisms (Hasher & Zacks, 1988; Lustig, Hasher, & Zacks, 2007). One other potential contributor to older adults' deficiencies in episodic memory is their relative inability to form and retrieve links among single bits of information, referred to as associative memory (Castel & Craik, 2003; Naveh-Benjamin, 2000; Naveh-Benjamin, Hussain, Guez, & Bar-On, 2003; Old & Naveh-Benjamin, 2008). Examples of associative memory include (but are not limited to) remembering who said what (source memory), order of information presentation, which items appeared together (item pairs), or whether something was seen or heard. Deficits in associative memory abilities make it difficult to create new associations between event information or units, thus limiting the ability to encode information effectively and later retrieve it (Chalfonte & Johnson, 1996).

In addition to the deficits often observed on associative memory tasks, older adults show a tendency to "falsely remember" information (Jacoby & Rhodes, 2006) and may, at times, be more captured by misleading information compared to younger individuals (Jacoby, Bishara, Hessels, & Toth, 2005). It has been proposed that this tendency to misremember or falsely remember may be due to an increased reliance on more automatic memory processes such as familiarity, in light of difficulties with more controlled memory processes (i.e., precise recollection;

Jacoby & Rhodes, 2006). While the reliance on familiarity and the ability to remember the “gist” can lead to accurate recall and create conditions that allow for more flexibility within memory and transfer of learning to novel situations (Koutstaal, 2006), it also often leads to higher occurrences of false remembering. For example, Jacoby (1999) had older and younger adults read a list of words one, two, or three times (thus increasing familiarity with those words when they were read multiple times). Participants then heard a separate list they were told to remember. During test, participants were told they would see words they had both read and heard, but only to respond to words that were heard. Interestingly, the increased repetition of the read words decreased younger adults’ false recognition, but increased older adults’ false recognition, indicating that older adults were relying more on familiarity of material during responding, possibly due to difficulties with exact recollection.

Although some degree of memory loss and memory changes may be inevitable with age, research is beginning to show that even in the types of memory most vulnerable to senescent changes, the ability to remember valuable, meaningful, and goal-relevant information may remain largely intact (Zacks & Hasher, 2006). As previously discussed, the socioemotional selectivity theory posits that older and younger adults have different motivations and goals concerning social interactions and emotional regulation, and this can have an impact on what older adults attend to and remember. Older adults have been shown to preferentially attend to positive compared with negative information (Isaacowitz, Wadlinger, Goren, & Wilson, 2006; Mather & Carstensen, 2003), and this differential allocation of attention can either enhance or decrease memory for emotional information. Thus, it is not surprising that older adults frequently remember a higher proportion of positive relative to negative information (i.e., they demonstrate a positivity bias), whereas younger adults either do not show this pattern or display a negativity bias in memory on both laboratory tasks (Charles, Mather, & Carstensen, 2003; Mather & Carstensen, 2005; Mather & Knight, 2005) as well as in their spontaneous autobiographical memories (Schlagman, Schulz, & Kvavilashvili, 2006; Tomaszczyk, Fernandes, & MacLeod, 2008). Consistent with the idea that emotional biases in memory may be a result of goal-directed processes (see Molden & Higgins, Chapter 20), older adults with the most pronounced positivity bias are those

who also score highest on tests of cognitive control capabilities (Mather & Carstensen, 2005).

Similar to emotional materials, information and scenarios that utilize more real-world, realistic, or relevant materials may serve to increase attention, motivation for remembering, and allow for the use of prior knowledge, thereby mitigating age-related memory impairments. Figure 33.7 displays the results of a study conducted by Castel (2005) that examined memory for prices of everyday grocery items. If the items were realistically priced, there were no age-related associative memory impairments for prices of grocery items, whereas large age-related decrements were present when older adults were asked to remember unrealistic prices. This finding highlights what a marked impact the utilization of meaningful, “real-world” materials can have on older adults’ performance on memory tasks. That is, when required to remember information that is consistent with prior knowledge, older adults can reduce their reliance on effortful, self-initiated processes (which may be detrimentally effected in aging), improving both encoding and retrieval memory operations (Castel, 2008; Craik & Bosman, 1992; McGillivray & Castel, 2010).

Hess and colleagues (Germain & Hess, 2007; Hess et al., 2001) have investigated the role of personal relevance and its impact on memory performance in older (and younger) adults. Hess et al. (2001) found that older adults were more accurate in their recollection of information related to a narrative describing an older target person (increased relevance), compared with one describing a younger target person, and this accuracy increased in situations

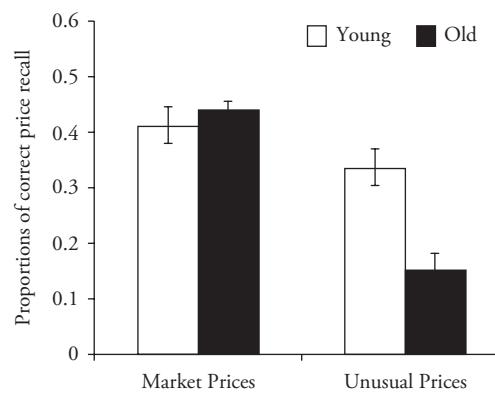


Fig. 33.7 The average proportion of correctly recalled prices by younger and older adults for the market value and unusually priced items. (From Castel, 2005. Copyright © 2005 by the American Psychological Association. Reproduced with permission.)

in which they were held accountable for their responses (increased motivation). Furthermore, older adults' memory benefitted to a greater extent from increasing motivation and relevance than did younger adults. Extending these findings, Germain and Hess (2007) demonstrated that increased relevance was strongly associated not only with memory performance but with more efficient processing, and that these effects were stronger within the older adult sample.

Motivation to remember and relevance are not always products of the to-be-remembered information but can reflect other situational variables. Adams and colleagues have investigated memory for stories, manipulating who participants (both younger and older women) were asked to retell a story to (an experimenter or a young child) (Adams, Smith, Pasupathi, & Vitolo, 2002). When the listener was an experimenter, younger adults recalled more propositional content than did older adults, but this age difference disappeared when the listener was a young child. Furthermore, when the listener was a child, both younger and older participants engaged in more elaborations and repetitions while retelling the story, but older adults were more adaptive in adjusting the complexity levels given the age of the listener. These findings underscore the importance of the context (in particular, social contexts) in which one is asked to recall information, and the degree to which differing context provides motivation to both younger and older adults.

Increasing relevance or importance of information can also serve to mitigate deficits in source memory (i.e., memory for information about the contextual details accompanying that event) so often observed among older adults. For example, no age-related differences were observed when older adults were asked to recognize whether a statement was true, false, or new (truth source), whereas large age-related differences were present when asked to identify the voice source (John or Mary said it) or whether it was a "new" statement (Rahhal, May, & Hasher, 2002). Similarly, older adults' memory performance equals younger adults' on source memory tasks when the to-be-remembered information has an emotionally relevant component (e.g., safety) (May, Rahhal, Berry, & Leighton, 2005). Lastly, recent research has shown that despite age-related memory declines, older adults are capable of remembering more important information just as well as younger adults, but this occurs at the expense of less important information (Castel, 2008; Castel, Benjamin,

Craik, & Watkins, 2002; Castel, McGillivray, & Friedman, 2011; Castel et al., in press).

Metacognition

Although this section is largely concerned with memory, it is also important to understand metacognitive processes and how these processes are affected by aging. Metacognition (or more specifically, metamemory) refers to one's awareness of his or her own memory and how it works. Metamemory includes, but is not limited to, beliefs about one's memory skills and task demands, insight into memory changes, feelings and emotions about one's memory, and knowledge of memory functioning (Dunlosky & Metcalfe, 2009). Beliefs that older and younger adults may have about their memory abilities, in turn, can influence expectations for memory performance, effort exerted during a memory task, and the degree to which one chooses to place himself or herself in demanding memory situations, and it can even influence one's actual performance (Dixon, Rust, Feltmate, & Kwong See, 2007; Lachman, 2006; Lachman & Andreoletti, 2006). Older adults are often very aware of deficits in memory performance (Hertzog & Hultsch, 2000; Levy & Leifheit-Limson, 2009), making the study of metamemory very important in terms of developing strategies to combat age-related memory decline.

Experimental studies of metamemory tasks often involve asking participants to make judgments of learning (or JOLs) about what or how much they will later remember (a form of metacognitive *monitoring*), or by asking participants what information they feel they need to restudy or study for shorter/longer periods of time (a form of metacognitive *control*). Investigations into the effects of age on these variable have been somewhat mixed. While some studies have found that old adults exhibit a larger pattern of overconfidence in their memory abilities compared with younger adults (i.e., there is a larger discrepancy between JOLs and actual memory performance; Bruce, Coyne, & Botwinick, 1982; Connor, Dunlosky, & Hertzog, 1997), other studies have found little to no age differences (Hines, Touron, & Hertzog, 2009; Lovelace & Marsh, 1985; Murphy, Sanders, Gabriesheski, & Schmitt, 1981), or more accurate performance by older adults (Hertzog, Dunlosky, Powell-Moman, & Kidder, 2002; Rast & Zimprich, 2009). In addition, recent work suggests that, relative to younger adults, older adults are also aware of how much information they have forgotten when learning and recalling lists of

items (Halamish, McGillivray, & Castel, 2011), suggesting that the monitoring of forgetting may be relatively intact in old age.

In regard to metacognitive control, Dunlosky and Connor (1997) observed that when older and younger adults were allowed to restudy words at their own pace, all participants spent more time studying items that they had been assigned lower JOLs (i.e., words they judged as more difficult to recall) compared with those words that had been given higher JOLs (i.e., judged as more likely to remember). However, younger adults exhibited this effect to a greater extent, indicating that age-related differences were present in the degree to which monitoring was used to effectively allocate study time. Dunlosky and Connor suggest that this difference in study-time allocation may even contribute to the lower overall memory performance in older adults. However, Dunlosky and Hertzog (1997) found that younger and older adults used a “functionally identical algorithm” in their selection of items for restudy, and both younger and older individuals adaptively selected to restudy the items they believed were not as well learned (Hines et al., 2009).

While the results surrounding metacognition and aging are somewhat mixed, it is encouraging that, at least under some conditions, monitoring and control over learning remains relatively intact throughout the life span. Even studies that have found sizable metacognitive deficits in older adults (e.g., Bunnell, Baken, & Richards-Ward, 1999) have also usually found that these deficits are less than those associated with actual memory ability. That is, metamemory abilities are likely better preserved in older adults than explicit memory abilities. This relative sparing suggests that older adults may be able to use metacognitive strategies to help overcome or compensate for age-related declines in memory performance.

Summary

Declines in older adults' memory abilities are perhaps one of the most widely documented findings within cognitive aging, and older adults frequently remark on their own difficulties with remembering. Older adults, more often than younger adults, remember less information overall, have difficulties forming associations between information, and are more likely to experience false or inaccurate memories. However, if the to-be-remembered information is more personally relevant, realistic (i.e., consistent with prior knowledge), valuable, or more

emotionally meaningful, age-related differences can be reduced. Research examining older adults' metacognitive abilities has yielded mixed results, with some studies documenting impairments, and others finding no age-related changes. Thus, there is at least some evidence suggesting the abilities to correctly predict one's memory abilities and monitor one's learning processes remain somewhat intact, at least compared to explicit memory abilities. This could be a result of lifelong experience “working” and learning to understand one's own dynamic memory capacities.

Expertise and Training

A majority of the theories and evidence discussed earlier in this chapter indicate that older adults' cognitive performance across several domains declines with age (e.g., Salthouse, 1985). But what role might expertise or training play in mitigating these effects? Many cultures consider old age to be associated with maturity and vast amounts of experience, as can be illustrated in the medical, musical, and even business fields (Krampe & Charness, 2006). Indeed, a majority of Fortune 500 CEOs range in age from their late 50s to early 60s. This begs the question whether older adults who are considered “experts” in a certain profession or skill are exempt from age-related declines and can function just as efficiently in their trade as their younger counterparts. What role does training have in maintaining expertise in older adulthood and how does it differ from the training of young adults? Finally, are the effects of training long lasting and differentially beneficial in older adulthood?

Expertise

The first, and arguably most important, question to consider is whether older adults who are classified as experts in a specific skill are exempt from age-related declines. The results are somewhat mixed, depending on the nature of the assessment given and how related it is to the mastered skill (Morrow et al., 2003). For example, Castel (2007) tested younger and older adults, as well as a group of retired older accountants and bookkeepers for their ability to recall object, numeric, and location information (e.g., 26 cherries in a bowl). The results revealed that the older adult experts performed just as well as the older controls in recollection of object information (skill unrelated), but those same experts outperformed younger adults (and older adults) in their memory for the numeric

information (skill related). This finding supports the notion that there are some basic limitations to expertise in old age, in the sense that mastery in one skill (recall of unrelated numbers) does not necessarily translate into high performance of another skill that is less related (recall of objects; see also Salthouse & Maurer, 1996).

However, expertise within some fields may serve to enhance certain cognitive capabilities, such that there may be some transfer effects into other domains (Chase & Ericsson, 1982; Krampe & Charness, 2006). To illustrate this point, Shimamura and colleagues (1995) examined the ability to recall prose information (pertaining to various topics) in younger, middle-aged, and older professors, as well as in college-educated younger and older adult “nonprofessor” controls. Older adult controls displayed deficits in recalling prose information relative to younger controls. However, processing and remembering dense passages is something that professors do frequently and is thus an area in which they could be considered experts. Among the groups of professors no age-related differences were found, despite the fact that the to-be-remembered material was not directly related to their fields of study. These results suggest there can be benefits for remaining highly cognitively active in old age, in that it may mitigate declines in certain memory abilities.

Similarly, Krampe and Ericsson (1996) suggest that lifelong experience and use of an acquired skill is sufficient to sustain lifelong expertise (Meinz, 2000; Meinz and Salthouse, 1998; Salthouse, 1991; Salthouse et al., 1990; but see Krampe, Engbert, & Kliegl, 2002; Krampe, Mayr, & Kliegl, 2005). This assumption was supported by Charness (1981a; 1981b) who found that although older adult chess players came up with fewer potential moves than their younger equivalents, the moves they selected were of equal quality. Thus, even older adults who are classified as experts are still susceptible to reduced cognitive resources and have to consider only the valuable or relevant information as opposed to every possible option. Indeed this “refinement” may also be associated with training or maintenance of expert skill levels, and many older adults claim that their practice is more “efficient” than when they were younger (Krampe, 1994; Krampe & Ericsson, 1996). This is consistent with the selective optimization with compensation model which claims that older adults can use novel or alternative means to counter losses in certain functions (Baltes & Baltes, 1990).

Training

Older adults, to some extent, can benefit from specific training designed to enhance or preserve cognitive abilities. A majority of studies have explored interventions to maintain, if not eliminate, age-related declines. For example, Willis et al. (2006) had older adults participate in a cognitive training intervention known as ACTIVE (Advanced Cognitive Training for Independent and Vital Elderly), which included memorial, inductive, and speed of processing training, and measured daily functions and cognitive abilities after an extensive delay (5 years). Although the training did not eliminate age-related functional declines in everyday activities, it did substantially slow their progression. Cavallini, Pagnin, and Vecchi (2003) illustrated similar findings by training working memory. Both younger and older adults benefited from the training, but younger adults' memory performance was still better than that of older adults. This study illustrates that although working memory does deteriorate in normal aging, older adults can still learn new information and strategies to counteract the decline (but see Dumitriu et al., 2010). Training can also lead to benefits in self-monitoring, making older adults more aware of what information they have not learned as well, which they should opt to study for longer periods of time (Dunlosky, Kubat-Silman, & Hertzog, 2003).

The long-lasting effects of training for older adults are comparable to those for younger adults, but daily use of the strategies learned is the best predictor of such benefits. For example, memory performance was similar to posttraining measures, given maintenance of practice for older adults after a 2-year delay (Bottiroli, Cavallini, & Vecchi, 2008). Derwinger, Neely, and Bäckman (2005) gave older adult participants either structured training or participant-generated mnemonic training. They found superior memory performance for the generated mnemonic group 8 months posttraining, even though the structured training group still showed a long-term benefit of training relative to controls (see also West, Bagwell, & Dark-Fruedeman, 2008). Benefits of training have also been demonstrated for shorter time scales. When using an incremented-difficulty approach (i.e., adding more and more intervening trials between test trials), older adults were able to correctly recollect information across increasing delays (Jennings & Jacoby, 2003). Overall, it appears as though cognitive training cannot completely eliminate declines in cognitive function (see

Hertzog, Kramer, Wilson, & Lindenberger, 2009), but it can be effective at slowing the rate of decline, especially if the training is incorporated into everyday life or the strategies used are self-generated, as the impact can be relatively long lasting.

Summary

Older adult experts are often exempt from age-related declines, but only for tasks that are related to the skill in which they acquired expertise. Even in light of declines, abilities that are frequently maintained or refined across the lifespan may allow older adults to continue to function optimally within skill-specific domains. While unable to completely stop or reverse age-related declines, cognitive training in older adulthood can slow declines via specific strategies designed to counteract specific detriments. In particular, self-generated techniques, and frequent use of learned strategies, make the benefits of training more robust and long lasting; however, the transfer of these skills to other domains is often limited.

Wisdom and Successful Aging

While it is clear that cognitive decline typically accompanies old age, many older adults are highly successful individuals who are high-functioning and are respected for their wisdom. For example, many CEOs, world leaders (or advisors), and deans of major universities are older adults who are recognized for their wisdom and expertise, and are entrusted with making important decisions and solving difficult problems (see also Salthouse, 2010). While the study and definition of wisdom is often elusive, most would conceptualize wisdom as expert knowledge or experiences that help inform future decision making and behavior (Baltes & Smith, 1990; but see Jeste et al., 2010). In addition, wisdom is often mentioned in the same breath as creativity and sometimes genius (see Sternberg, 1985, also Simonton, Chapter 25). Thus, while the concept of wisdom is still elusive in terms of a precise definition and components, it is clear that we can recognize the usefulness of wisdom, and we often turn to people rich in wisdom for guidance and trust their judgment. While various forms of cognitive processes seem to slow or are impaired in old age, it is widely believed that wisdom often increases with age and life experience. In fact, as discussed by Goldberg (2006) in his book *The Wisdom Paradox*, people associate wisdom with advancing age (Orwoll & Perlmuter, 1990) and also regard wisdom as one

of the most desirable traits (Heckhausen, Dixon, & Baltes, 1989), clearly demonstrating there are some positive aspects to arriving at old age.

In an attempt to measure the contribution of age to social wisdom, Grossmann et al. (2010) had participants read stories about intergroup and interpersonal conflicts, and they were then asked to predict the end result of these conflicts. Compared to young and middle-aged adults, the older adults used higher order reasoning schemes that emphasize the need for taking multiple perspectives, allowing for compromise, and the recognition of the limits of knowledge (Grossmann et al., 2010). This finding suggests that in contrast to other types of reasoning that are typically measured in the lab and are found to decline with age (see Salthouse, 2000), some forms of social reasoning may actually improve with age and life experience.

Research has also shown that creative pursuits are influenced by age. Lehman (1953) outlined how production of superior lyrical poetry and music typically shows a peak between the ages of 25 to 29 but also again at the age range of 80 to 84 (see also Simonton, 1998). In addition, the cognitive processes that lead to creative output at an early age may be altered or controlled by completely different mechanisms than those that contribute to creative output in old age. This is clearly an avenue for future research, but what is apparent is that the odds of producing great work is related to the number of attempts, suggesting that perseverance and wisdom may enhance creativity in older adults. In addition, people often change roles due to lifelong experience, such as taking on new jobs, teaching roles, or advisor positions, or simply by taking different perspectives due to expertise and knowledge. The use of creativity and wisdom in later life can then be linked directly to successful aging (Adams-Price, 1998). For example, while Michelangelo and Einstein had some of their most productive years at an early age, their wisdom was then often called upon later in life to provide advice and insight regarding important decisions and events. Nora Ochs recently became the oldest person ever to finish college when, at age 95, she completed a degree in history and graduated on the same day as her 21-year-old granddaughter, demonstrating that perhaps the key to creativity and enjoyment in old age is engaging in active pursuits.

According to theorists Rowe and Kahn (1998), successful aging can be defined as a combination of several key elements. These include an absence of diseases and disabilities; dealing with changes in

control, bereavement, and social support; maintaining high levels of physical and cognitive abilities; and preserving social and productive activities. From a more behaviorist perspective, toward the end of his career and well into old age himself, B.F. Skinner wrote a book on how to enjoy old age (Skinner & Vaughn, 1983). Although he outlined the numerous limiting factors associated with aging, he also focused on the many positive aspects of aging and the need to selectively focus on certain goals (c.f. Baltes & Baltes, 1990), as well as the need to have an optimistic perspective regarding life and development. While creativity, wisdom, and successful aging are central themes in life-span development, there is a clear need to better understand how specific cognitive processes and perspectives contribute to successful aging.

Conclusions and Future Directions

Although some declines in cognitive capabilities may be inevitable with age, a growing body of research has begun to emphasize the sizable impact that factors such as goals, motivation, prior knowledge, and experience have on older adults' performance across a variety of domains. In addition, given the broad and diverse changes that can accompany aging, future research needs to examine how thinking is impaired and enhanced in older adulthood by considering the effects of the factors mentioned earlier, as well as culture, wisdom, and expertise. It is not enough to document impairments, as research has identified many areas in which older adults show qualitatively different approaches to problem solving, incorporate emotional content when making decisions, and are often more experienced than younger adults. Thus, a more comprehensive and multidimensional approach to the study of age-related changes is warranted, one that considers the dynamic interaction of motivational, emotional, and biological changes and the impact these factors can have on cognitive processes (see also Hess, 2005). In addition, the manner in which older adults can judiciously determine what information is important, use that information to facilitate memory and decision making, and then communicate important information to others in an efficient manner, is an interesting avenue for future research (see also Castel, McGillivray, & Friedman, 2011). Lastly, the use of technology has greatly changed how people can access information when making decisions and when trying to remember information. Today, more and more older adults are using the Internet and hand-held devices (Charness & Boot, 2009). The

access and use of technology, and how this modifies thinking for older adults (e.g., Small, Moody, Siddarth, & Bookheimer, 2009) is an important direction for future research.

References

- Adams, C., Smith, M. C., Pasupathi, M., & Vitolo, L. (2002). Social context effects on story recall in older and younger women: Does the listener make a difference? *Journal of Gerontology: Psychological Sciences*, 57, P28–40.
- Adams-Price, C. E. (1998). *Creativity and successful aging*. New York: Springer.
- Allaire, J. C., & Marsiske, M. (2002). Well- and ill-defined measures of everyday cognition: Relationship to older adults' intellectual ability and functional status. *Psychology and Aging*, 17, 101–115.
- American Association of Retired Persons. (1996). *Telemarketing fraud and older Americans: An AARP survey*. Washington, D.C.: Author.
- Anderson, N. D., Craik, F. I. M., & Naveh-Benjamin, M. (1998). The attentional demands of encoding and retrieval in younger and older adults: Evidence from divided attention costs. *Psychology and Aging*, 13, 405–423.
- Artistico, D., Orom, H., Cervone, D., Krauss, S., & Houston, E. (2010). Everyday challenges in context: The influence of contextual factors on everyday problem solving among young, middle-aged, and older adults. *Experimental Aging Research*, 36, 230–247.
- Babcock, R. L. (2002). Analysis of age differences in types of errors on the Raven's advanced progressive matrices. *Intelligence*, 30, 485–503.
- Baltes, P. B., & Baltes, M. M. (1990). Psychological perspectives on successful aging: The model of selective optimization with compensation. In P. B. Baltes & M. M. Baltes (Eds.), *Successful aging: Perspectives from the behavioral sciences* (pp. 1–34). New York: Cambridge University Press.
- Baltes, P., & Smith, J., (1990). Toward a psychology of wisdom and its ontogenesis. In R. Sternberg (Ed.), *Wisdom: Its nature, origins, and development* (pp. 87–120). New York: Cambridge University Press.
- Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, 50, 7–15.
- Berg, C. A., Meegan, S. P., & Klaczynski, P. (1999). Age and experiential differences in strategy generation and information requests for solving everyday problems. *International Journal of Behavioral Development*, 23, 615–639.
- Blanchard-Fields, F., Chen, Y., & Norris, L. (1997). Everyday problem solving across the adult life span: Influence of domain specificity and cognitive appraisal. *Psychology and Aging*, 12, 684–693.
- Blanchard-Fields, F., Jahnke, H. C., & Camp, C. (1995). Age differences in problem-solving style: The role of emotional salience. *Psychology and Aging*, 10, 173–180.
- Blanchard-Fields, F., Mienaltowski, A., & Seay, R. B. (2007). Age differences in everyday problem-solving effectiveness: Older adults select more effective strategies for interpersonal problems. *Journal of Gerontology: Psychological Sciences*, 62, P61–64.
- Blanchard-Fields, F., Stein, R., & Watson, T. L. (2004). Age differences in emotion-regulation strategies in handling

- everyday problems. *Journal of Gerontology: Psychological Sciences*, 59, P261–269.
- Blieszner, R., Willis, S. L., & Baltes P. B. (1981). Training research in aging on the fluid ability of inductive reasoning. *Journal of Applied Developmental Psychology*, 2, 247–265.
- Boone, K. B., Ghaffarian, S., Lesser, I. M., & Hill-Gutierrez, E. (1993). Wisconsin Card Sorting Test performance in healthy, older adults: Relationship to age, sex, education, and IQ. *Journal of Clinical Psychology*, 49, 54–60.
- Bottiroli, S., Cavallini, E., & Vecchi, T. (2008). Long-term effects of memory training in the elderly: A longitudinal study. *Archives of Gerontology and Geriatrics*, 47, 277–289.
- Braver, T. S., & Barch, D. M. (2002). A theory of cognitive control, aging cognition, and neuromodulation. *Neuroscience and Biobehavioral Reviews*, 26, 809–817.
- Bruce, P. R., Coyne, A. C., & Botwinick, J. (1982). Adult age differences in metamemory. *Journal of Gerontology*, 37, 354–357.
- Bunnell, J. K., Baken, D. M., & Richards-Ward, L. A. (1999). The effect of age on metamemory for working memory. *New Zealand Journal of Psychology*, 28, 23–29.
- Burton, C. L., Strauss, E., Hultsch, D. F., & Hunter, M. A. (2006). Cognitive functioning and everyday problem solving in older adults. *The Clinical Neuropsychologist*, 20, 432–452.
- Cabeza, R. (2001). Functional neuroimaging of cognitive aging. In R. Cabeza & A. Kingston (Eds.), *Handbook of functional neuroimaging of cognition* (pp. 331–377). Cambridge, MA: MIT Press.
- Carstensen, L. L. (1992). Social and emotional patterns in adulthood: Support for socioemotional selectivity theory. *Psychology and Aging*, 7, 331–338.
- Carstensen, L. L. (1995). Evidence for a life-span theory of socioemotional selectivity. *Current Directions in Psychological Science*, 4, 151–156.
- Carstensen, L. L., Gross, J., & Fung, H. (1997). The social context of emotion. In M. P. Lawton & K. W. Schaie (Eds.), *Annual review of geriatrics and gerontology* (pp. 325–352). New York: Springer.
- Carstensen, L. L., Pasupathi, M., Mayr, U., & Nesselroade, J. R. (2000). Emotional experience in everyday life across the adult life span. *Journal of Personality and Social Psychology*, 79, 644–655.
- Castel, A. D. (2005). Memory for grocery prices in younger and older adults: The role of schematic support. *Psychology and Aging*, 20, 718–721.
- Castel, A. D. (2007). Aging and memory for numerical information: The role of specificity and expertise in associative memory. *Journal of Gerontology: Psychological Sciences*, 62, 194–196.
- Castel, A. D. (2008). The adaptive and strategic use of memory by older adults: Evaluative processing and value-directed remembering. In A. S. Benjamin & B. H. Ross (Eds.), *The psychology of learning and motivation* (Vol. 48, pp. 225–270). London: Academic Press.
- Castel, A. D., Benjamin, A. S., Craik, F. I. M., & Watkins, M. J. (2002). The effects of aging on selectivity and control in short-term recall. *Memory and Cognition*, 30, 1078–1085.
- Castel, A. D., & Craik, F. I. M. (2003). The effects of aging and divided attention on memory for item and associative information. *Psychology and Aging*, 18, 873–885.
- Castel, A. D., Humphreys, K. L., Lee, S. S., Galvan, A., Balota, D. A., McCabe, D. P. (in press). The development of memory efficiency and value-directed remembering across the lifespan: A cross-sectional study of memory and selectivity. *Developmental Psychology*.
- Castel, A. D., McGillivray, S., & Friedman, M. C. (2011). Metamemory and memory efficiency in older adults: Learning about the benefits of priority processing and value-directed remembering. In M. Naveh-Benjamin & N. Ohta (Eds.), *Memory and aging: Current issues and future directions* (pp. 243–268). Philadelphia, PA: Psychology Press.
- Cavallini, E., Pagnin, A., & Vecchi, T. (2003). Aging and everyday memory: The beneficial effect of memory training. *Archives of Gerontology and Geriatrics*, 37, 241–257.
- Chalfonte, B. L., & Johnson, M. K. (1996). Feature memory and binding in young and older adults. *Memory and Cognition*, 24, 403–416.
- Charles, S. T., Mather, M., & Carstensen, L. L. (2003). Aging and emotional memory: The forgettable nature of negative images for older adults. *Journal of Experimental Psychology: General*, 132, 310–324.
- Charness, N. (1981a). Aging and skilled problem solving. *Journal of Experimental Psychology: General*, 110, 21–38.
- Charness, N. (1981b). Search in chess: Age and skill differences. *Journal of Experimental Psychology: Human perception and Performance*, 7, 467–476.
- Charness, N., & Boot, W. R. (2009). Aging and information technology use: Potential and barriers. *Current Directions in Psychological Science* 18, 253–258.
- Chase, W. G., & Ericsson, K. A. (1982). Skill and working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 16, pp. 1–58). San Diego, CA: Academic Press.
- Cheng, S., & Strough, J. (2004). A comparison of collaborative and individual everyday problem solving in younger and older adults. *The International Journal of Aging and Human Development*, 58, 167–195.
- Coats, A. H., & Blanchard-Fields, F. (2008). Emotion regulation in interpersonal problems: The role of cognitive-emotional complexity, emotion regulation goals, and expressivity. *Psychology and Aging*, 23, 39–51.
- Connor, L. T., Dunlosky, J., & Hertzog, C. (1997). Age-related differences in absolute but not relative metamemory accuracy. *Psychology and Aging*, 12, 50–71.
- Craik, F. I. M., & Bosman, B. A. (1992). Age-related changes in memory and learning. In H. Bouma & J. A. M. Graafmans (Eds.), *Gerontechnology* (pp. 79–92). Amsterdam, Netherlands: IOS Press.
- Craik, F. I. M., & Byrd, M. (1982). Aging and cognitive deficits: The role of attentional resources. In F. I. M. Craik & S. E. Trehub (Eds.), *Aging and cognitive processes* (pp. 191–211). New York: Plenum.
- Craik, F. I. M., & Salthouse, T. A. (2008). *Handbook of aging and cognition* (3rd ed.). Mahwah, NJ: Erlbaum.
- Crawford, S., & Channon, S. (2002). Dissociation between performance on abstract tests of executive function and problem solving in real-life-type situations in normal aging. *Aging and Mental Health*, 6, 12–21.
- Darowski, E. S., Helder, E., Zacks, R. T., Hasher, L., & Hambrick, D. Z. (2008). Age-related differences in cognition: The role of distraction control. *Neuropsychology*, 22, 638–644.
- Davis, H. P., Klebe, K. J., Bever, B., & Spring, A. (1998). The effect of age on the learning of a nondeclarative category classification task. *Experimental Aging Research*, 24, 24–41.
- Denburg, N. L., Cole, C. A., Hernandez, M., Yamada, T. H., Tranel, D., Bechara, A., & Wallace, R. B. (2007). The orbitofrontal

- cortex, real-world decision making, and normal aging. *Annals of the New York Academy of Sciences*, 1121, 480–498.
- Denburg, N. L., Tranel, D., & Bechara, A. (2005). The ability to decide advantageously declines prematurely in some normal older persons. *Neuropsychologia*, 43, 1099–1106.
- Denney, N. W., & Palmer, A. M. (1981). Adult age differences on traditional and practical problem-solving measures. *Journal of Gerontology*, 36, 323–328.
- Denney, N. W., & Pearce, K. A. (1989). A developmental study of practical problem solving in adults. *Psychology and Aging*, 4, 438–442.
- Denney, N. W., Pearce, K. A., & Palmer, A. M. (1982). A developmental study of adults' performance on traditional and practical problem-solving tasks. *Experimental Aging Research*, 8, 115–118.
- Derwinger, A., Neely, A. S., & Bäckman, L. (2005). Design your own memory strategies! Self-generated strategy training versus mnemonic training in old age: An 8-month follow-up. *Neuropsychological Rehabilitation*, 15, 37–54.
- Diehl, M., Willis, S. L., & Schaie, K. W. (1995). Everyday problem solving in older adults: Observational assessment and cognitive correlates. *Psychology and Aging*, 10, 478–491.
- Dixon, R. A., Rust, T. B., Feltmate, S. E., & Kwong See, S. (2007). Memory and aging: Selected research directions and application issues. *Canadian Psychology*, 48, 67–76.
- Dumitriu, D., Hao, J., Hara, Y., Kaufmann, J., Janssen, W. G. M., Lou, W., et al. (2010). Selective changes in thin spine density and morphology in monkey prefrontal cortex correlate with aging-related cognitive impairment. *The Journal of Neuroscience*, 30, 7507–7515.
- Dunlosky, J., & Connor, L. T. (1997). Age differences in the allocation of study time account for age differences in memory performance. *Memory and Cognition*, 25, 691–700.
- Dunlosky, J., & Hertzog, C. (1997). Older and younger adults use a functionally identical algorithm to select items for restudy during multitrial learning. *Journal of Gerontology: Psychological Sciences*, 52, P178–186.
- Dunlosky, J., Kubat-Silman, A., & Hertzog, C. (2003). Training metacognitive skills improves older adults' associative learning. *Psychology and Aging*, 18, 340–345.
- Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Thousand Oaks, CA: Sage.
- Fein, G., McGillivray, S., & Finn, P. (2007). Older adults make less advantageous decisions than younger adults: Cognitive and psychological correlates. *Journal of the International Neuropsychological Society*, 13, 480–489.
- Filoteo, V. J., & Maddox, T. W. (2004). A quantitative model-based approach to examining aging effects on information-integration category learning. *Psychology and Aging*, 19, 171–182.
- Finucane, M. L., Slovic, P., Hibbard, J. H., Peters, E., Mertz, C. K., & MacGregor, D. G. (2002). Aging and decision-making competence: An analysis of comprehension and consistency skills in older versus younger adults considering health-plan options. *Journal of Behavioral Decision Making*, 15, 141–164.
- Fleischman, D. A., Wilson, R. S., Gabrieli, J. D., Bienias, J. L., & Bennett, D. A. (2004). A longitudinal study of implicit and explicit memory in old persons. *Psychology and Aging*, 19, 617–625.
- Fredrickson, B. L., & Carstensen, L. L. (1990). Choosing social partners: How old age and anticipated endings make people more selective. *Psychology and Aging*, 5, 335–347.
- Friedman, M. C., & Castel, A. D. (2010, April). *Memory, aging, and interference in a value-based encoding task*. Poster presented at the Cognitive Aging Conference 2010, Atlanta, GA.
- Friedman, M. C., Castel, A. D., McGillivray, S., & Flores, C. C. (2010, April). *Associative memory for money and faces in young and old adults*. Poster presented at the Cognitive Aging Conference 2010, Atlanta, GA.
- Fung, H. H., Carstensen, L. L., & Lutz, A. M. (1999). Influence of time on social preferences: Implications for life-span development. *Psychology and Aging*, 14, 595–604.
- Germain, C. M., & Hess, T. M. (2007). Motivational influences on controlled processing: Moderating distractibility in older adults. *Neuropsychology, Development, and Cognition. Section B, Aging Neuropsychology and Cognition*, 14, 462–486.
- Goldberg, E. (2006). *The wisdom paradox: How your mind can grow stronger as your brain grows older*. New York: Gotham Books.
- Grossmann, I., Na, J., Varnum, M. E. W., Park, D., Kitayama, S., & Nisbett, R. (2010). Reasoning about social conflicts improves into old age. *Proceedings of the National Academy of Science USA*, 107, 7246–7250.
- Halamish, V., McGillivray, S., & Castel, A. D. (2011). Impaired memory, intact metacognition: Monitoring one's own forgetting by younger and older adults. *Psychology and Aging*, 26, 631–635.
- Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and new view. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 22, pp. 193–225). New York: Academic Press.
- Heckhausen, J. (1999). *Developmental regulation in adulthood: Age normative and sociostructural constraints as adaptive challenges*. New York: Cambridge University Press.
- Heckhausen, J., Dixon, R., & Baltes, P. (1989). Gains and losses in development throughout adulthood as perceived by different adult age groups. *Developmental Psychology*, 25, 109–121.
- Heckhausen, J., & Schulz, R. (1995). A life-span theory of control. *Psychological Review*, 102, 284–304.
- Henninger, D. E., Madden, D. J., & Huettel, S. A. (2010). Processing speed and memory mediate age-related differences in decision-making. *Psychology and Aging*, 25, 262–270.
- Hertzog, C., Dunlosky, J., Powell-Moman, A., & Kidder, D. P. (2002). Aging and monitoring associative learning: Is monitoring accuracy spared or impaired? *Psychology and Aging*, 17, 209–225.
- Hertzog, C., & Hultsch, D. F. (2000). Metacognition in adulthood and old age. In F. I. M. Craik (Ed.), *The handbook of aging and cognition* (2nd ed., pp. 417–466). Mahwah, NJ: Erlbaum.
- Hertzog, C., Kramer, A. F., Wilson, R. S., & Lindenberger, U. (2009). Enriched effects on adult cognitive development: Can the functional capacity of older adults be preserved and enhanced? *Psychological Sciences in the Public Interest*, 9, 1–65.
- Hess, T. M. (2005). Memory and aging in context. *Psychological Bulletin*, 131, 383–406.
- Hess, T. M., Rosenberg, D. C., & Waters, S. J. (2001). Motivation and representational processes in adulthood: The effects of social accountability and information relevance. *Psychology and Aging*, 16, 629–642.

- Hess, T. M., & Slaughter, S. J. (1986). Specific exemplar retention and prototype abstraction in young and old adults. *Psychology and Aging, 1*, 202–207.
- Hicks-Patrick, J., & Strough, J. (2004). Everyday problem solving: Experience, strategies, and behavioral intentions. *Journal of Adult Development, 11*, 9–18.
- Hines, J. C., Touron, D. R., & Hertzog, C. (2009). Metacognitive influences on study time allocation in an associative recognition task: An analysis of adult age differences. *Psychology and Aging, 24*, 462–475.
- Hoppmann, C. A., Coats, A. H., & Blanchard-Fields, F. (2008). Goals and everyday problem solving: Examining the link between age-related goals and problem-solving strategy use. *Aging, Neuropsychology, and Cognition, 15*, 401–423.
- Isaacowitz, D. M., Wadlinger, H. A., Goren, D., & Wilson, H. R. (2006). Selective preference in visual fixation away from negative images in old age? An eye-tracking study. *Psychology and Aging, 21*, 40–48.
- Jacoby, L. L. (1999). Ironic effects of repetition: Measuring age-related differences in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 3–22.
- Jacoby, L. L., Bishara, A. J., Hessel, S., & Toth, J. P. (2005). Aging, subjective experience, and cognitive control: Dramatic false remembering by older adults. *Journal of Experimental Psychology: General, 134*, 131–148.
- Jacoby, L. L., & Rhodes, M. (2006). False remembering in the aged. *Current Directions in Psychological Science, 15*, 49–53.
- Jamieson, B., & Rogers, W. (2000). Age-related effects of blocked and random practice schedules on learning a new technology. *Journal of Gerontology: Psychological Sciences, 55B*, P343–P353.
- Jennings, J. M., & Jacoby, L. L. (2003). Improving memory in older adults: Training recollection. *Neuropsychological Rehabilitation, 13*, 417–440.
- Jeste, D. V., Ardel, M., Blazer, D., Kraemer, H. C., Vaillant, G., & Meeks, T. M. (2010). Expert consensus on characteristics of wisdom: A Delphi method study. *The Gerontologist, 50*, 668–680.
- Kausler, D. H. (1994). *Learning and memory in normal aging*. San Diego, CA: Academic Press.
- Kim, S., Goldstein, D., Hasher, L., & Zacks, R. T. (2005). Framing effects in younger and older adults. *Journal of Gerontology: Psychological Sciences, 60*, P215–218.
- Kim, S., & Hasher, L. (2005). The attraction effect in decision making: Superior performance by older adults. *Quarterly Journal of Experimental Psychology, 58*, 120–133.
- Kim, S., Healey, M. K., Goldstein, D., Hasher, L., & Wiprzycka, U. J. (2008). Age differences in choice satisfaction: A positivity effect in decision making. *Psychology and Aging, 23*, 33–38.
- Kimbler, K. J., & Margrett, J. A. (2009). Older adults' interactive behaviors during collaboration on everyday problems: Linking process and outcome. *International Journal of Behavioral Development, 33*, 531–542.
- Kornell, N., Castel, A. D., Eich, T. S., & Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and Aging, 25*, 498–503.
- Koutstaal, W. (2006). Flexible remembering. *Psychonomic Bulletin and Review, 13*, 84–91.
- Kovalchik, S., Camerer, C. F., Grether, D. M., Plott, C. R., & Allman, J. M. (2005). Aging and decision making: A comparison between neurologically healthy elderly and young individuals. *Journal of Economic Behavior and Organization, 58*, 79–94.
- Krampe, R. T. (1994). *Maintaining excellence: Cognitive-motor performance in pianists differing in age and skill level*. Berlin, Germany: Edition Sigma.
- Krampe, R. T., & Charness, N. (2006). Aging and experience. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 723–742). New York: Cambridge University Press.
- Krampe, R. T., Engbert, R., & Kliegl, R. (2002). The effects of expertise and age on rhythm production: Adaptations to timing and sequencing constraints. *Brain and Cognition, 48*, 179–194.
- Krampe, R. T., & Ericsson, K. A. (1996). Maintaining excellence: Deliberate practice and elite performance in young and older pianists. *Journal of Experimental Psychology: General, 125*, 331–359.
- Krampe, R. T., Mayr, U., & Kliegl, R. (2005). Timing, sequencing, and executive control in repetitive movement production. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 206–233.
- Krawczyk, D. C., Morrison, R. G., Viskontas, I., Holyoak, K. J., Chow, T. W., Mendez, M. F., et al. (2008). Distraction during relational reasoning: The role of prefrontal cortex in interference control. *Neuropsychologia, 46*, 2020–2032.
- Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity? *Intelligence, 14*, 389–433.
- Lachman, M. E. (2006). Perceived control over aging-related declines: Adaptive beliefs and behaviors. *Current Directions in Psychological Science, 15*, 282–286.
- Lachman, M. E., & Andreoletti, C. (2006). Strategy use mediates the relationship between control beliefs and memory performance for middle-aged and older adults. *Journal of Gerontology: Psychological Sciences, 61*, P88–94.
- Laver, G. D. (2009). Adult aging effects on semantic and episodic priming in word recognition. *Psychology and Aging, 24*, 28–39.
- Lavoie, D. J., & Cobia, D. J. (2007). Recollecting, recognizing, and other acts of remembering: An overview of human memory. *Journal of Neurologic Physical Therapy, 31*, 135–144.
- Lehman, H. C. (1953). *Age and achievement*. Princeton, NJ: Princeton University Press.
- Levy, B. R., & Leifheit-Limson, E. (2009). The stereotype-matching effect: Greater influence on functioning when age stereotypes correspond to outcomes. *Psychology and Aging, 24*, 230–233.
- Levy-Cushman, J., & Abeles, N. (1998). Memory complaints in the able elderly. *Clinical Gerontologist, 19*, 3–24.
- Light, L. L., & Singh, A. (1987). Implicit and explicit memory in young and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 531–541.
- Lockenhoff, C. E., & Carstensen, L. L. (2007). Aging, emotion, and health-related decision strategies: Motivational manipulations can reduce age differences. *Psychology and Aging, 22*, 134–146.
- Lovelace, E. A., & Marsh, G. R. (1985). Prediction and evaluation of memory performance by young and old adults. *Journal of Gerontology, 40*, 192–197.
- Lustig, C., Hasher, L., & Zacks, R. T. (2007). Inhibitory deficit theory: Recent developments in a "new view." In

- D. S. Gorfein & C. M. MacLeod (Eds.), *The place of inhibition in cognition* (pp. 145–162). Washington, DC: American Psychological Association.
- MacPherson, S. E., Phillips, L. H., & Della Sala, S. (2002). Age, executive function, and social decision making: A dorsolateral prefrontal theory of cognitive aging. *Psychology and Aging, 17*, 598–609.
- Marsiske, M., & Willis, S. L. (1995). Dimensionality of everyday problem solving in older adults. *Psychology and Aging, 10*, 269–283.
- Mata, R. (2007). Understanding the aging decision maker. *Human Development, 50*, 359–366.
- Mata, R., & Nunes, L. (2010). When less is enough: Cognitive aging, information search, and decision quality in consumer choice. *Psychology and Aging, 25*, 289–298.
- Mata, R., Schoeler, L. J., & Rieskamp, J. (2007). The aging decision maker: Cognitive aging and the adaptive selection of decision strategies. *Psychology and Aging, 22*, 796–810.
- Mata, R., von Helversen, B., & Rieskamp, J. (2010). Learning to choose: Cognitive aging and strategy selection learning in decision making. *Psychology and Aging, 25*, 299–309.
- Mather, M., & Carstensen, L. L. (2003). Aging and attentional biases for emotional faces. *Psychological Science, 14*, 409–415.
- Mather, M., & Carstensen, L. L. (2005). Aging and motivated cognition: The positivity effect in attention and memory. *Trends in Cognitive Sciences, 9*, 496–502.
- Mather, M., & Johnson, M. K. (2003). Affective review and schema reliance in memory in older and younger adults. *American Journal of Psychology, 116*, 169–189.
- Mather, M., & Knight, M. (2005). Goal-directed memory: The role of cognitive control in older adults' emotional memory. *Psychology and Aging, 20*, 554–570.
- May, C. P., Rahhal, T., Berry, E. M., & Leighton, E. A. (2005). Aging, source memory, and emotion. *Psychology and Aging, 20*, 571–578.
- McCabe, D. P., Roediger, H. L., McDaniel, M. A., Balota, D. A., & Hambrick, D. Z. (2010). The relationship between working memory capacity and executive functioning: Evidence for a common executive attention construct. *Neuropsychology, 24*, 222–243.
- McGillivray, S., & Castel, A. D. (2010). Memory for age-face associations: The role of generation and schematic support. *Psychology and Aging, 25*, 822–832.
- Meinz, E. J. (2000). Experience-based attenuation of age-related differences in music cognition tasks. *Psychology and Aging, 15*, 297–312.
- Meinz, E. J., & Salthouse, T. A. (1998). The effects of age and experience on memory for visually presented music. *Journals of Gerontology: Psychological Sciences, 53B*, 60–69.
- Morrow, D. G., Ridolfo, H. E., Menard, W. E., Sanborn, A., Stine-Morrow, E. A., Magnor, C., et al. (2003). Environmental support promotes expertise-based mitigation of age differences on pilot communication task. *Psychology and Aging, 18*, 268–284.
- Murphy, M. D., Sanders, R. E., Gabriesheski, A. S., & Schmitt, F. A. (1981). Metamemory in the aged. *Journal of Gerontology, 36*, 185–193.
- Mutter, S. A., & Plumlee, L. F. (2009). Aging and integration of contingency evidence in causal judgment. *Psychology and Aging, 24*, 916–926.
- Naveh-Benjamin, M. (2000). Adult age differences in memory performance: Tests of an associative deficit hypothesis. *Journal of Experimental Psychology: Learning Memory and Cognition, 26*, 1170–1187.
- Naveh-Benjamin, M., Hussain, Z., Guez, J., & Bar-On, M. (2003). Adult age differences in episodic memory: Further support for an associative-deficit hypothesis. *Journal of Experimental Psychology: Learning Memory and Cognition, 29*, 826–837.
- Nilsson, L. G. (2003). Memory function in normal aging. *Acta Neurologica Scandinavica, 179*, 7–13.
- Old, S. R., & Naveh-Benjamin, M. (2008). Differential effects of age on item and associative measures of memory: A meta-analysis. *Psychology and Aging, 23*, 104–118.
- Orwoll, L., & Perlmuter, M. (1990). The study of wise persons: Integrating a personality perspective. In R. Sternberg (Ed.), *Wisdom: Its nature, origins, and development* (pp. 160–180). New York: Cambridge University Press.
- Pachur, T., Mata, R., & Schoeler, L. J. (2009). Cognitive aging and the adaptive use of recognition in decision making. *Psychology and Aging, 24*, 901–915.
- Ostreicher, M. L., Moses, S. N., Rosenbaum, R. S., & Ryan, J. D. (2010). Prior experience supports new learning of relations in aging. *Journal of Gerontology: Psychological Sciences, 65B*, 32–41.
- Park, D. C., Smith, A. D., Dudley, W. N., & Lafronza, V. N. (1989). Effects of age and a divided attention task presented during encoding and retrieval on memory. *Journal of Experimental Psychology: Learning Memory and Cognition, 15*, 1185–1191.
- Peters, E., Finucane, M. L., MacGregor, D. G., & Slovic, P. (2000). The bearable lightness of aging: Judgment and decision processes in older adults. In P. C. Stern & L. L. Carstensen (Eds.), *The aging mind: Opportunities in cognitive research* (pp. 144–165). Washington, DC: National Academies Press.
- Peters, E., Hess, T. M., Västfjäll, D., & Auman, C. (2007). Adult age differences in dual information processes: Implications for the role of affective and deliberative processes in older adults' decision making. *Perspectives on Psychological Science, 2*, 1–23.
- Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353–363.
- Racine, C. A., Barch, D. M., Braver, T. S., & Noelle, D. C. (2006). The effect of age on rule-based category learning. *Aging, Neuropsychology, and Cognition, 13*, 411–434.
- Rahhal, T. A., May, C. P., & Hasher, L. (2002). Truth and character: Sources that older adults can remember. *Psychological Science, 13*, 101–105.
- Rast, P., & Zimprich, D. (2009). Age differences in the underconfidence-with-practice effect. *Experimental Aging Research, 35*, 400–431.
- Raz, N., Gunning, F. M., Head, D., Dupuis, J. H., McQuain, J., Briggs, S. D., et al. (1997). Selective aging of the human cerebral cortex observed in vivo: Differential vulnerability of the prefrontal gray matter. *Cerebral Cortex, 7*, 268–282.
- Reed, A. E., Mikels, J. A., & Simon, K. I. (2008). Older adults prefer less choice than younger adults. *Psychology and Aging, 23*, 671–675.
- Rhodes, M. G. (2004). Age-related differences in performance on the Wisconsin Card Sorting Test: A meta-analytic review. *Psychology and Aging, 19*, 482–494.

- Ridderinkhof, K. R., Span, M. M., & van der Molen, M. W. (2002). Perseverative behavior and adaptive control in older adults: Performance monitoring, rule induction, and set shifting. *Brain and Cognition*, 49, 382–401.
- Riediger, M., Li, S. C., & Lindenberger, U. (2006). Selection, optimization, and compensation as developmental mechanisms of adaptive resource allocation: Review and preview. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology and aging* (6th ed., pp. 289–313). Amsterdam, Netherlands: Elsevier.
- Rowe, J. W., & Kahn, R. L. (1998). *Successful aging*. New York: Dell Publishing.
- Ryan, J. D., Moses, S. N., & Villate, C. (2008). Impaired relational organization of propositions, but intact transitive inference, in aging: Implications for understanding underlying neural integrity. *Neuropsychologia*, 47, 338–353.
- Salthouse, T. A. (1985). Speed of behavior and its implications for cognition. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 400–426). New York: Van Nostrand Reinhold.
- Salthouse, T. A. (1991). Age and experience effects on the interpretation of orthographic drawings of three-dimensional objects. *Psychology and Aging*, 6, 426–433.
- Salthouse, T. A. (1993). Influence of working memory on adult age differences in matrix reasoning. *British Journal of Psychology*, 84, 171–199.
- Salthouse, T. A. (1994). The nature of the influence of speed on adult age differences in cognition. *Developmental Psychology*, 30, 240–259.
- Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, 103, 403–428.
- Salthouse, T. A. (2000). Aging and measures of processing speed. *Biological Psychology*, 54, 35–54.
- Salthouse, T. A. (2005). Effects of aging on reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 589–605). New York: Cambridge University Press.
- Salthouse, T. A. (2010). *Major issues in cognitive aging*. New York: Oxford University Press.
- Salthouse, T. A., Babcock, R. L., Skovronek, E., Mitchell, D. R. D., & Palmon, R. (1990). Age and experience effects in spatial visualization. *Developmental Psychology*, 26, 128–136.
- Salthouse, T. A., & Maurer, T. J. (1996). Aging, job performance, and career development. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (4th ed., pp. 353–364). New York: Academic Press.
- Salthouse, T. A., & Skovronek, E. (1992). Within-context assessment of age differences in working memory. *Journal of Gerontology: Psychological Sciences*, 47, 110–120.
- Samanez-Larkin, G. R., Gibbs, S. E., Khanna, K., Nielsen, L., Carstensen, L. L., & Knutson, B. (2007). Anticipation of monetary gain but not loss in healthy older adults. *Nature Neuroscience*, 10, 787–791.
- Schlagman, S., Schulz, J., & Kvavilashvili, L. (2006). A content analysis of involuntary autobiographical memories: Examining the positivity effect in old age. *Memory*, 14, 161–175.
- Shimamura, A. P., Berry, J. M., Mangels, J. A., Rusting, C. L., & Jurica, P. J. (1995). Memory and cognitive abilities in academic professors: Evidence for successful aging. *Psychological Science*, 6, 271–277.
- Simonton, D. K. (1998). Career paths and creative lives: A theoretical perspective on late life potential. In C. E. Adam-Price (Ed.), *Creativity and successful aging* (pp. 3–18). New York: Springer Publication Company.
- Skinner, B. F., & Vaughn, M. E. (1983). *Enjoy old age: A practical guide*. New York: Norton.
- Small, G. W., Moody, T. D., Siddarth, P., & Bookheimer, S. Y. (2009). Your brain on Google: Patterns of cerebral activation during Internet searching. *American Journal of Geriatric Psychiatry*, 17, 116–126.
- Soederberg-Miller, L. M., & West, R. L. (2010). The effects of age, control beliefs, and feedback on self-regulation of reading and problem solving. *Experimental Aging Research*, 36, 40–63.
- Sternberg, R. (1985). Implicit theories of intelligence, creativity and wisdom. *Journal of Personality and Social Psychology*, 49, 607–627.
- Strough, J., Cheng, S., & Swenson, L. M. (2002). Preferences for collaborative and individual everyday problem solving in later adulthood. *International Journal of Behavioral Development*, 26, 26–35.
- Strough, J., Hicks-Patrick, J., Swenson, L. M., Cheng, S., & Barnes, K. A. (2003). Collaborative everyday problem solving: Interpersonal relationships and problem dimensions. *The International Journal of Aging and Human Development*, 56, 43–66.
- Strough, J., McFall, J. P., Flinn, J. A., & Schuller, K. L. (2008). Collaborative everyday problem solving among same-gender friends in early and later adulthood. *Psychology and Aging*, 23, 517–530.
- Strough, J., Mehta, C. M., McFall, J. P., & Schuller, K. L. (2008). Are older adults less subject to the sunk-cost fallacy than younger adults? *Psychological Science*, 19, 650–652.
- Tanis, B., Wood, S., Hanoch, Y., & Rice, T. (2009). Aging and choice: Applications to Medicare Part D. *Journal of Judgment and Decision Making*, 4, 92–101.
- Thornton, W. J., & Dumke, H. A. (2005). Age differences in everyday problem-solving and decision-making effectiveness: A meta-analytic review. *Psychology and Aging*, 20, 85–99.
- Tomaszyk, J. C., Fernandes, M. A., & MacLeod, C. M. (2008). Personal relevance modulates the positivity bias in recall of emotional pictures in older adults. *Psychonomic Bulletin and Review*, 15, 191–196.
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211, 453–458.
- Verhaeghen, P. (2003). Aging and vocabulary scores: A meta-analysis. *Psychology and Aging*, 18, 332–339.
- Verhaeghen, P., & Salthouse, T. A. (1997). Meta-analyses of age-cognition relations in adulthood: Estimates of linear and nonlinear age effects and structural models. *Psychological Bulletin*, 122, 231–249.
- Viskontas, I. V., Morrison, R. G., Holyoak, K. J., Hummel, J. E., & Knowlton, B. J. (2004). Relational integration, inhibition, and analogical reasoning in older adults. *Psychology and Aging*, 19, 581–591.
- Viskontas, I. V., Holyoak, K. J., & Knowlton, B. J. (2005). Relational integration in older adults. *Thinking and Reasoning*, 11, 390–410.
- Watson, T. L., & Blanchard-Fields, F. (1998). Thinking with your head and your heart: Age differences in everyday problem-solving strategy preferences. *Aging, Neuropsychology, and Cognition*, 5, 225–240.

- West, R. L. (1996). An application of prefrontal cortex function theory to cognitive aging. *Psychological Bulletin*, 120, 272–292.
- West, R. L., Bagwell, D. K., & Dark-Fruedeman, A. (2008). Self-efficacy and memory aging: The impact of a memory intervention based on self-efficacy. *Aging, Neuropsychology, and Cognition*, 15, 302–329.
- Willis, S. L., Tennstedt, S. L., Marsiske, M., Ball, K., Elias, J., Koepke, K. M., et al. (2006). Long-term effects of cognitive training on everyday functional outcomes in older adults. *Journal of the American Medical Association*, 296, 2805–2814.
- Wood, S., Busemeyer, J., Koling, A., Cox, C. R., & Davis, H. (2005). Older adults as adaptive decision makers: Evidence from the Iowa Gambling Task. *Psychology and Aging*, 20, 220–225.
- Zacks, R. T., & Hasher, L. (2006). Aging and long-term memory: Deficits are not inevitable. In E. Bialystok & F. I. M. Craik (Eds.), *Lifespan cognition: Mechanisms of change* (pp. 162–177). New York: Oxford University Press.

The Cognitive Neuroscience of Thought Disorder in Schizophrenia

Peter Bachman and Tyrone D. Cannon

Abstract

The term *thought disorder* most commonly refers to a constellation of impairments in communication manifested by individuals suffering from schizophrenia. Although diverse in nature, these symptoms are thought to result from the influence of one or a small number of cognitive abnormalities that affect how individuals with psychotic disorders process information. We discuss the phenomenology of thought disorder and the candidate cognitive mechanisms that may play a role in its expression. Among these, impaired executive functioning—possibly through its interaction with semantic memory—shows the greatest promise in accounting for the phenomena that comprise thought disorder. Additionally, we review a prominent model of executive control of ongoing behavior and discuss links to psychosis symptoms. We then outline a set of neurophysiological abnormalities associated with schizophrenia, or with latent genetic risk for developing the disorder, and consider how these factors may contribute to the expression of disordered thinking.

Key Words: formal thought disorder, psychosis, working memory, speech production, context, endophenotype, discourse coherence, prefrontal, temporal

Introduction

During the course of two assessment interviews, the following standardized, open-ended questions were asked of two research subjects:

[Examiner] “Can you explain the proverb, ‘Speech is the picture of the mind?’”

[Study participant] “You see the world through speech. Like my grandfather used to speak to me of Alaskans and Alsations and blood getting thicker and thinner in the Eskimo. He was against the Kents in England. I can’t smoke a Kent cigarette to this day” (p. 44, Harrow & Quinlan, 1985).

[Examiner] “Why should people pay taxes?”

[Study participant] “Taxes is an obligation as citizens of our country. To our nation, to this country, the United States. As a citizen, I think we have

an obligation. I think that’s carried to an extreme. Within reason, taxes within reason. Taxation, we have representation, so therefore we have taxation. For we formed our constitution, it was taxation without representation is treason” (p. 263, Johnston & Holzman, 1979).

Reading these two responses evokes the feeling that something is not quite right; they seem somehow disordered. For instance, the sentences comprising the first response convey messages that are only very indirectly related to each other, and the types of associations are inconsistent (e.g., concept of “speech” apparently linking the first two sentences, and the idea of participant’s grandfather—the object—linking the second and third). Consequently, rather than a linear progression toward a predictable goal, the response follows a

tangent that deviates dramatically from the content requested by the examiner. The reply to the second examiner's question does not follow such a rapidly digressing course, but instead it seems to fixate on an idea, or perhaps phrase ("taxation without representation"), indirectly related to the content of the question, seeming to repeat and elaborate on that idea or phrase in a circumstantial manner, without offering additional ideas. Both examples involve disrupted communication, which we interpret to reflect disordered underlying thought processes.

Apart from describing how these statements are disordered, understanding why the speakers produced them in such a manner is a daunting yet intellectually stimulating task. The process of comprehending and generating speech integrated into an ongoing conversation involves numerous interrelated cognitive (Levelt, 1989) and neural mechanisms (Price, 2010), any or all of which could contribute to characteristic abnormalities in speech comprehension or production. Additionally, the thought disorder apparent in these examples tends to occur within the context of a more extensive psychopathology (Andreasen & Grove, 1986), including diagnoses as diverse as schizophrenia, mood disorders, certain personality disorders, and autism (American Psychiatric Association, 2000; Andreasen & Grove, 1986). In fact, the patient quoted in the first reply was diagnosed with schizophrenia, the condition perhaps most closely associated with the presence of thought disorder (e.g., Bleuler, 1911/1950). The second quote, however, was provided by an individual who was not herself diagnosed with a psychiatric disorder but has a daughter who was diagnosed with schizoaffective disorder (a condition thought to be closely related to schizophrenia; American Psychiatric Association, 2000), highlighting the role of heritable factors contributing to significant symptom expression, even in the absence of meeting diagnostic criteria for a syndrome.

Despite the prevalence of thought disorder across diagnostic populations, we are not aware of systematic efforts to study the pathology of thought disorder across diverse conditions, looking for common disease mechanisms. Rather, most investigators have chosen to study thought disorder strictly within the context of a particular disease "entity" such as schizophrenia. However, this restriction does not necessarily go very far in simplifying the search for the ultimate causes of thought disorder. The diagnoses most closely associated with the presence of thought

disorder are among the most complex, most etiologically obscure conditions encountered in modern medicine (e.g., Abrahams & Geschwind, 2010; Schulze, 2010; Tandon, Keshavan, & Nasrallah, 2008). Considering schizophrenia, the plethora of difficulties many patients face—including a range of degraded information processing capabilities, presence of debilitating symptoms (including hallucinations and delusions), medication side effects, substantial social and occupational difficulties—defies models of etiology based on a single underlying deficit. In recognition of this complexity, psychopathology researchers have begun to dissect disorders whose manifestation coincides with the presentation of thought disorder into more fundamental phenotypic traits that ultimately participate in symptom formation. Recent work adopting this approach has demonstrated, for instance, that certain neurocognitive disruptions in schizophrenia are associated with genetic vulnerability to the illness, other traits are associated with disease expression, and some may be attributable to any number of interactions between these two (e.g., Cannon, et al., 2000; Cannon, Thompson, et al., 2002; Jaaro-Peled, Ayhan, Petnikov, & Sawa, 2010; McGrath et al., 2009; Rutten & Mill, 2009).

This more complex, integrative view of the pathology of cognitive impairment, in the context of particular disease states (Cannon & Rosso, 2002), may very well be necessary to our ultimate elucidation of the set of neurobiological and cognitive conditions that may participate in the expression of thought disorder through their collective action. In this chapter, we focus first on the role of thought disorder in psychopathology, highlighting descriptive approaches. We then shift to discussion of a prominent model of speech production and how it may map on to brain function (see Morrison & Knowlton, Chapter 6), as well as a prominent model of disordered thinking in schizophrenia, in order to help us identify neurocognitive mechanisms likely disrupted in individuals displaying thought disorder. Finally, we attempt to integrate findings from distinct levels of analysis (e.g., behavioral and molecular genetics, structural and functional neuroanatomy, experimental psychopathology), in order to characterize diverse aspects of psychiatric disorder as traits specific to disease expression, which generally tend to implicate the brain's temporal lobe structures, and traits specific to genetic vulnerability, which tend to involve more frontal lobe-mediated functions.

Defining Thought Disorder

Perhaps the most common usage of the term *thought disorder* is as shorthand for *formal thought disorder*, which refers to a taxonomy of symptoms involving abnormal speech (Andreasen, 1979, 1982). Historically, formal thought disorder has typically been conceptualized as the product of loosened associations leading to a loss of continuity between the ordered elements in the thought processes that precipitate a spoken utterance (Maher, 1991). The “formal” distinction in particular harkens back to the notion that pathologies of thought can be characterized as disorders of thought content or as disorders of thought form. In the schizophrenia literature, disordered content referred primarily to delusions, objectively false and often bizarre beliefs held with a high level of conviction (i.e., the patient maintains the belief in the face of counter-evidence; American Psychiatric Association, 2000). A common example of a delusion is the belief that people in the patient’s environment intend to harm the patient (i.e., a paranoid delusion). On the other hand, disorders of thought form were believed to involve a disorganization of underlying thought processes, reflected in abnormal speech, such as the passages quoted at the outset of this chapter. Although this form-versus-content distinction is no longer commonly used in clinical practice, factor analytic studies of symptom prevalence in schizophrenia (e.g., Liddle, 1987, Peralta, Cuesta, & Farre, 1997) have generally supported the distinction. Ratings of formal thought disorder load together with ratings of disorganized behavior on a factor distinct from delusions and hallucinations, suggesting that thought disorder symptoms indeed reflect a disorganization of ideational elements not necessarily specific to articulated speech.

Several examples of phenomena characterized as formal thought disorder are pertinent to our discussion of ideational disorganization. One category of speech abnormality is the generation of neologisms, or novel words formed by the unique integration of parts of other words. A neologism would therefore be conceptualized as the loosening of normal associative relationships between individual word parts (perhaps at the level of grammatical encoding, discussed later in this chapter). Johnston and Holzman (1979) quoted one patient as responding to an examiner’s request to define the word *remorse* by replying, “*Moisterous, being moistful*” (p. 100), combining legal word parts to form lexically invalid words.

Similarly, an affected individual’s speech may be characterized by lexically valid but unrelated words strung together to make an unintelligible statement—a loosening of associations between words. An example of this type of disordered comment is, “If things turn by rotation of agriculture or levels or timed in regard to everything . . .” (Maher, 1966, p. 395). In its extreme form clinicians sometimes refer to this type of disorganization as “word salad,” indicating that the words do not blend together in any semantically meaningful way.

Formal thought disorder may also manifest itself in an abrupt shift between indirectly related topics, representing a loosened association between ideas or clauses, within or between sentences. For example, when one patient with formal thought disorder was asked to interpret the proverb “One swallow does not make the summer,” he replied, “When swallow in your throat like a key it comes out, but not as scissors. A robin, too, it means spring” (quoted in Harrow & Quinlan, 1985, p. 429). In this instance, the patient seems to have switched from employing one meaning of the word *swallow* (i.e., the verb) to an alternate meaning (i.e., the type of bird), and then articulating a concept (i.e., “robin”) semantically related to the alternate meaning.

Perhaps even more salient in this last example than the abrupt shift is that the patient’s response seems only very tangentially consistent with the interviewer’s question. In fact, disordered speech can involve statements that are overly vague or overly concrete, or otherwise do not seem congruent with the semantic or interpersonal demands implied by the comment or question posed by the other participant in the conversation.

In contrast to the taxonomy of speech abnormalities described earlier, which follows from application of the ideational confusion definition, Andreasen (1986) developed a descriptive system of assessing thought disorder, intended to eschew traditional assumptions about the pathology resulting in disordered speech. In all, Andreasen identified 18 classes of speech abnormality, which she showed could be clustered using factor analysis into clinically meaningful categories (Andreasen & Grove, 1986).

Indeed, five of the 18 types of abnormality clustered together to comprise a “loose associations” dimension that seemed to indicate the overall level of behavioral disorganization shown by patients with psychotic disorders (Andreasen & Grove, 1986). Another distinction appeared between what the authors characterized as aspects of “positive” and

“negative” thought disorder (analogous to but not isomorphic with the positive-negative schizophrenia symptom distinction), with the former involving aspects of loosened associations (e.g., derailment) in combination with a significant level of rapidity and volume of speech (sometimes referred to as “pressure” of speech), and the latter involving speech that is impoverished in terms of average number and length utterance and/or impoverished with respect to ideational content (as was the case in the second quote cited in the Introduction). Interestingly, this positive versus negative dimension effectively discriminated thought-disordered patients with mania (in the context of bipolar disorder) from thought-disordered patients with schizophrenia.

An area of long-running controversy in the study of formal thought disorder is whether thought disorder is ultimately a disorder of thought itself or a disorder of overt speech. Specifically, rather than considering this markedly disrupted ability to communicate a speech production problem, we infer that the locus of pathology lies in the thought processes underlying the intentional production of speech (see Chaika, 1982, for additional discussion of this inference).

Unfortunately, as discussed in depth by Critchley (1964) and Maher (1991), these thought processes themselves are not directly observable. Therefore, measurement of any putative disruption must necessarily occur indirectly, usually with the assumption that the psychomotor transformation from a thought to a spoken utterance occurs with a normal range of fidelity. It is certainly worthwhile to consider whether this assumption is warranted, as doing so will provide a heuristic for considering the stages of communication that must be disrupted in order for particular “thought disorder” phenomena to be manifest.

Levelt's Model of Normal Speech Production

To provide an organizing framework for our consideration of how specific cognitive impairments may contribute to the expression of formal thought disorder, we turn first to a model of normal speech production.

As shown in Figure 34.1, Levelt and colleagues' (Indefrey & Levelt, 2004; Levelt, 1989, 1999; Levelt, Roelofs, & Meyer, 1999) model involves a serial process by which a message intended for

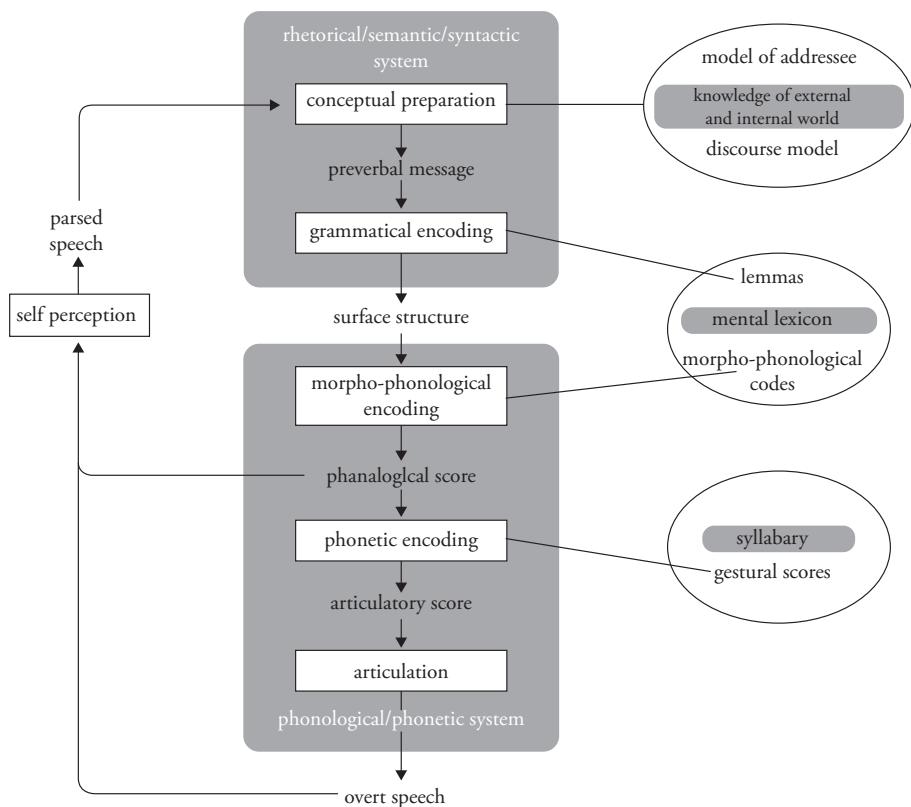


Fig. 34.1 Levelt's model of normal speech production (from Levelt, 1999).

communication moves through a succession of stages, each of which plays a unique role in transforming the message into an articulated sound wave. The first set of stages along this speech production sequence comprises what Levelt refers to as a “rhetorical/semantic/syntactic system,” responsible for filtering a given communicative intention through the speaker’s model of how the listener will perceive and understand the message, which can be influenced by the speaker’s mental model of the listener. This system also sequences ideas in a logical order and places that sequence in a propositional format (specific to linguistic expression), which includes the selection of lexical concepts, in turn triggering the retrieval of appropriate lemmas (abstract, canonical forms of words) from the mental lexicon. The retrieval of the appropriate lemmas from the mental lexicon engages the syntactic construction of the message, as lemmas must agree syntactically with each other and with the overall communicative intent of the speaker.

This retrieval of lemmas from the mental lexicon, which also entails retrieval of each lemma’s inherent morpho-phonological code, serves as a transition out of the “rhetorical/semantic/syntactic system” and into the “phonological/phonetic system.” Indeed, the lemmas’ store in the mental lexicon represents the basic stage at which semantic and phonological information is bound together.

Accordingly, the phonological codes associated with the lemmas’ various morphemes combine according to the predetermined sequence to form the syllabic structure of the message. This is a relative process, the product of which does not necessarily respect the boundaries of the superordinate lemmas. Next, during the process of phonetic encoding, the phonological score (i.e., the accumulation of the phonological syllables) retrieves a gestural, or articulatory score, completing the process by which a fully formed syntactic and phonological message retrieves an appropriate articulatory motor plan. Subsequently, articulation, the generation of overt speech, is the physical realization of the selected motor plan.

The production of overt speech, however, does not represent the final stage in Levelt’s model of speech production. In fact, the model also includes a feedback loop by which the speaker can perceive and monitor his or her own speech for errors or external interference, reengaging the model at the level of conceptual preparation in order to make appropriate corrections if necessary.

As Ditman and Kuperberg (2010) discuss, speech production models such as this one must account not only for the selection of each subsequent word but also for the maintenance of larger discourse coherence, a logical consistency between clauses contributing to the buildup of a situation model, or a gestalt meaning the speaker is attempting to communicate. When discourse coherence is successfully maintained, referents are unambiguous (e.g., pronouns substituting for objects), and there is a logical continuity between statements heard by the speaker and statements made in response (see also Garrod & Pickering, 2004).

In fact, there is a wide range of temporal scales and complex linguistic structures across which logical consistency must be maintained. Although feed-forward models of language production such as Levelt’s tend not to emphasize its influence, the role of context can be considered at each of these levels. In this sense, context could be defined as any influence other than the lexico-semantic information built into previously spoken statements. How contextual influences may affect subsequent speech will be considered at length later.

On a neural level, Indefrey and Levelt (1999) describe the functioning of Levelt’s model as implemented in a primarily left hemisphere-lateralized cortical network. They propose that the initial process of conceptual preparation occurs in a range of heteromodal and cortical association areas (typically areas specific to the stimulus leading in to the pre-set message formulation; Indefrey & Levelt, 2004), the activity of which converges with the selection of a lexical concept occurring in the left middle temporal gyrus (Indefrey & Levelt, 1999) and the pars orbitalis and pars opercularis section of the inferior frontal gyrus, medial prefrontal cortex, posterior inferior parietal cortex, and parahippocampal and posterior cingulate gyri (Price, 2010). Subsequently, Wernicke’s area (roughly the temporal-parietal junction, including the supramarginal gyrus and the posterior planum temporale) and more anterior sections of the pars opercularis segment of the inferior frontal gyrus (Price, 2010) are activated by the retrieval of phonological codes associated with retrieval and sequencing of lexical concepts, followed by activation of Broca’s area (posterior left inferior frontal cortex and anterior insula) and the left midsuperior temporal lobe, the sites at which phonological encoding continues independent of lexical information. Broca’s area then remains active and is joined by activation in other supplementary

motor areas, left putamen, and cerebellum during the process of articulation (reviewed by Price, 2010). Indefrey and Levelt (1999) further specify that self-monitoring, whether occurring covertly or overtly, activates regions of superior temporal lobe, as well as supplementary motor areas related to articulation. In addition, the most abstract levels of monitoring for discourse coherence and congruence with speaker's goals and models of listener's understanding are almost certain to engage prefrontal areas (especially dorsolateral prefrontal cortex) and superior temporal gyrus (e.g., Sassa et al., 2007).

Thought Disorder or Speech Disorder?

Referring to Level's model of normal speech production (Fig. 34.1), we can consider each of the putative processing stages and attempt to infer what the observable product of "lesioning" each in isolation would sound like. Let us first examine the processing stage most closely affiliated with the actual act of speaking, the process of physical articulation. If an intact would-be utterance moves into the stage of physical articulation, only to be compromised, one might expect the output to contain the intended words, encoded grammatically, but spoken in a manner that systematically distorts the articulatory score of the phrase. Such a spoken product would not resemble formally disordered thought, but instead the product of conditions such as dysarthria or speech apraxia (Dronkers, Redfern, & Knight, 1999), two disorders familiar to neurologists and speech pathologists.

Similarly, if the lesion underlying formal thought disorder involved the process of phonetic encoding, one would expect spoken output to resemble speech characteristic of what Dodd and Crosbie (2010) refer to as "speech programming deficit disorder," in which speech is produced fluently, but the distorted phonological score would yield speech devoid of normal patterns of pitch and syllabification—perhaps sounding severely slurred—in the absence of dysarthria or speech apraxia.

The immediately preceding stage, morpho-phonological encoding, the first point in the process at which a word's phonological code is processed independently of semantic content, has also been shown to be compromised in isolation, specifically in patients suffering from an anomic aphasia, or a word-finding deficit (Indefrey & Levelt, 1999). Typically, such patients describe a sense of frustration over feeling that they have particular verbal concepts in mind but cannot retrieve the phonological code; that is, they cannot think of how to say the corresponding

word. Certainly this condition is debilitating, but apart from the superficial similarity with thought blocking (sometimes considered a feature of negative formal thought disorder but not a construct included in Andreasen's and Grove's system), anomic aphasia does not resemble formal thought disorder.

Finally, having ruled out deficits in the stages of speech production comprising Levelt's (1989, 1999) phonological/phonetic system, we work backward to the stage at which lemmas are selected and retrieved from the mental lexicon, initiating the grammatical encoding process. Agrammatic patients, such as patients suffering from Broca's aphasia, are characterized by speech in which words are selected and ordered appropriately, but the particular form of each word is not adjusted to accommodate the grammatical demands of nearby words or the phrase as a whole (e.g., verbs are not conjugated correctly; Indefrey & Levelt, 2004). Although, as apparent in the quotes at the beginning of the chapter, patients with formal thought disorder make grammatical mistakes in their speech, it is not clear that they make such mistakes more frequently than non-thought-disordered individuals do.

Evidence does exist (Andreasen & Grove, 1986; Berenbaum & Barch, 1995), however, that patients with formal thought disorder show a small, but significant level of word substitution and approximation, which is the predictable consequence of faulty retrieval of lemmas from the mental lexicon, the initial process occurring under the heading of grammatical encoding. We therefore conclude that we can rule out lesion to all processing stages occurring after lemma retrieval, up to and including the articulation of overt speech. The etiology of formal thought disorder must therefore exist somewhere along the way through the rhetorical/semantic/syntactic system (including application of a mental model of the listener and conceptual preparation) and/or in the self-monitoring feedback loop. Although where one draws the line between "thought" and "speech" is somewhat of a philosophical issue, we propose that all processes comprising these two suspect components certainly warrant the label "thought," justifying the term *formal thought disorder*, rather than *speech disorder*.

Overview of Cognitive Models of Thought Disorder

The first major psychological discussion of the pathology of thought disorder was provided by Eugen Bleuler, a Swiss psychiatrist and theorist contemporary

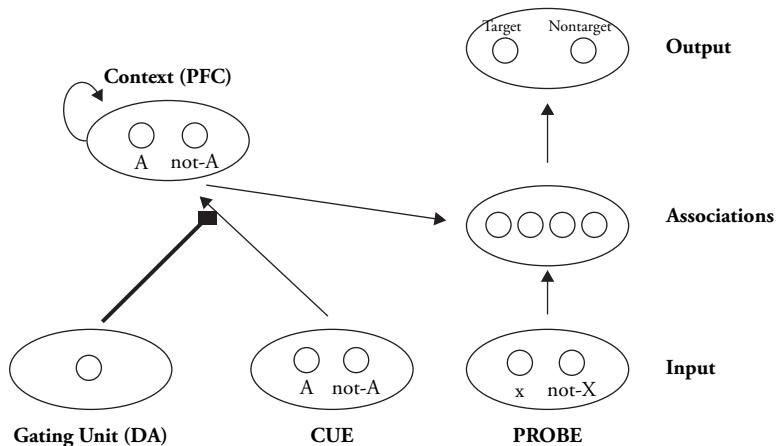


Fig. 34.2 Cohen, Braver, and colleagues' model of information processing disruption in schizophrenia (from Braver, Barch, & Cohen, 1999). PFC, prefrontal cortex.

to both Sigmund Freud, founder of modern clinical psychology, and Wilhelm Wundt, often cited as the founder of modern experimental psychology. Based on his observation of patients with psychotic disorders, such as schizophrenia, Bleuler (1911/1950) argued that the etiology of psychosis involves a fundamental “loosening of associations” between ideational elements, which results in a conceptual disorganization that manifests itself in disordered speech (in addition to other symptoms), an idea preserved almost exactly in its original form in more contemporary definitions of thought disorder, as discussed earlier.

Furthermore, Bleuler's conceptualization of the pathology of psychosis is analogous to more contemporary cognitive explanations of disordered information processing (e.g., Andreasen et al., 1999; Goldman-Rakic, 1995; Oltmanns & Neale, 1975; Silverstein, Kovacs, Corry, & Valone, 2000), as he proposed that a critical parameter of a fundamental *cognitive* mechanism is abnormal and that the consequences of this single cognitive “lesion” account for the diverse phenomena observed in the behavior of many psychiatric patients. One contemporary cognitive model—perhaps the dominant psychological model currently informing cognitive neuroscience research on schizophrenia—holds that schizophrenia patients' abnormal prefrontal functioning curtails normal integration of thought and behavior, as indicated by the inability to use contextual information in the efficient guidance of ongoing, goal-directed behavior.

Guided Activation Model

Jonathan Cohen, Earl Miller, and their colleagues (Braver, Barch, & Cohen, 1999; Braver et al., 2001;

Cohen, Barch, Carter, & Servan-Schreiber, 1999; Cohen & Servan-Schreiber, 1992; Miller & Cohen, 2001) have proposed a heuristic model of prefrontal control of ongoing behavior that has been applied extensively to characterize cognitive impairment in schizophrenia (see Fig. 34.2). By this logic, information processing deficits observed in schizophrenia patients result from a disturbance in the interaction between a prefrontal cortical module specialized for the representation, active maintenance, and updating of information regarding goal states, and a more posterior system responsible for the generation of sensory representations and the storage of learned behavioral contingencies. An individual's repertoire of stimulus-response associations must be directly accessible to the behavioral selection process that bridges the gap between the encoding of a stimulus and the execution of a response. Accordingly, the existence of a pathway allowing interaction between these stored behavioral contingencies (i.e., long-term memories, including motor plans) and the goal representation module allows this type of context information to influence the selection and execution of ongoing behavioral plans, ideally biasing behavior in a goal-appropriate manner (Braver et al., 1999). Context information, such as goal state, therefore mediates the selection of learned associations, which would otherwise be dictated by environmental stimuli. Cohen, Miller, and colleagues (Miller & Cohen, 2001) place particular emphasis on the idea that the prefrontal cortex representation module pushes competing, mutually inhibitory neural assemblies closer to activation, and in the case of a representation of behavior, closer to execution, but it does not necessarily select them outright. Goals,

and other context information, therefore guide the activation of behaviors, making them more or less likely to occur, given present circumstances.

To serve this function, context information must be represented and maintained in a manner that leaves it both buffered against interference (from task-irrelevant stimuli) and available to be updated as required by changing task demands. In an extreme example, this system must be capable of exercising cognitive control: utilizing information from a previous stimulus to favor the processing of relevant information and to suppress processing of irrelevant information, and then selecting appropriate goal-directed behavior, even in the face of competition from more salient behavioral responses.

Citing computational evidence (Braver et al., 1999), the investigators argue that variable efficiency of the interaction between the context-processing module and the learned behavioral contingencies module could in fact account for schizophrenia patients' reduced sensitivity to context information. They elaborate (Braver et al., 1999) that a gating mechanism must exist between the two modules, allowing context information to be encoded and maintained without interference from irrelevant perceptual information under certain circumstances; and at other times, making context information available for updating or to influence activation in the association storage module. Disrupted information processing in schizophrenia is therefore the consequence of failure of this "gate" between the association storage module and the context-processing module to properly open and close, degrading the ability of goal-related information to bias response selection.

In support of their model, Cohen, Braver, and colleagues (Cohen et al., 1999) have presented data from schizophrenia patients and controls performing three tasks: a single-trial version of the Stroop task, a lexical disambiguation task, and a "one-back" continuous performance task requiring subjects to continuously match each stimulus with the stimulus presented immediately prior (Cohen & Servan-Schreiber, 1992). In each task, the difficulty of maintaining context information and using it to select appropriate behavior was manipulated by varying the length of time during which context information must be maintained prior to response selection, as well as the salience of task-appropriate responses relative to task-inappropriate responses (i.e., the demand for cognitive control during the behavioral selection process). The investigators

argued that, overall, schizophrenia patients display a differential insensitivity to context information. This insensitivity interacted with variable information maintenance demands in two out of three experiments (Cohen et al., 1999). Additionally, the investigators reported a significant negative correlation between context sensitivity and severity of disorganization symptoms (including formal thought disorder) among schizophrenia patients, suggesting that the ability to effectively and flexibly bind ideational elements to an appropriate context underlies both the production of organized speech and successful performance on these context-heavy tasks.

Cohen, Braver, and colleagues have also offered evidence supporting a different but related aspect of their model: the assertion that context information is actively maintained, updated, and buffered against interference in the prefrontal cortex (chiefly, dorsolateral prefrontal areas; Barch et al., 1997; Braver et al., 1997; Braver, Reynolds, & Donaldson, 2003; Cohen et al., 1997). Additionally, Miller and Cohen (2001; see also Duncan, 2001; Kane & Engle, 2002) have reviewed evidence that the prefrontal cortex also modulates activity in regions of the brain associated with modality-specific buffers, specifically by holding long-term memories at a high level of activation.

On a functional anatomical level, these findings support Miller and Cohen's conception of the prefrontal cortex as a pan-modal modulator of modality-specific information processing centers, facilitating the selection of goal-appropriate behavior. Incorporating an additional level of analysis, Cohen, Braver, and colleagues cite evidence (Braver et al., 1999) that phasic dopamine activity modulates the gate between the prefrontal context-processing module and the individual's repertoire of learned behavioral contingencies. More recent work has indeed supported the role of phasic dopamine signaling in the updating and stabilization of representations; however, this work has pointed more reliably to dopamine signaling in the dorsal basal ganglia (reviewed by Simpson, Kellendonk, & Kandel, 2010) and how it may affect prefrontal-striatal circuits responsible for selecting and maintaining information in an active state (e.g., Landu, Lal, O'Neil, Baker, & Jagust, 2009; McNabb & Klingberg, 2008). Although further clarification is needed, the importance of dopamine signaling in modulating these prefrontal-striatal circuits is certainly evident.

In addition to being able to account qualitatively for the cognitive deficits the model was designed to simulate (Braver et al., 1999), the proposition that schizophrenia patients fail to appropriately use context information to guide ongoing behavior in goal-directed manner certainly has face validity. One might argue, however, that any behavior judged to be abnormal, or more specifically, deficient with respect to a given goal state, could be explained by a failure of this context-processing mechanism. The concept most critical to differentiating such general failures of goal-directed behavior from those reflecting disruptions of context processing is cognitive control. Specifically, patients will perform a given task correctly when the correct behavioral response is somehow most salient or dominant with respect to other potential responses; in this case, the representation of context and the prepotency of the correct response are redundant mechanisms. When the correct response is less salient, or less "prepotent" than an incorrect, distracter response, patients will tend to choose the distracter. Nonpsychotic subjects, conversely, will be more capable of using representations of context to inhibit the prepotent distracter and select the appropriate, less salient behavioral response; that is, they will be more capable of exercising cognitive control.

This focus on cognitive control therefore represents a critical step in the development of this model and allows it to benefit from additional specification regarding the neural correlates of cognitive control (e.g., Braver et al., 2003) and both neural and cognitive mechanisms underlying recognition and resolution of response conflict (Botvinick et al., 2001). Along these same lines, demonstration of how the heuristic guided activation model may subsume fully elaborated computational models of specific cognitive tasks, such as the Context Maintenance and Retrieval model of free recall performance in verbal memory tasks (Polyn, Norman, & Kahana, 2009), will aid in the generation of additional, testable hypotheses.

An issue awaiting resolution also bears mentioning: the mechanism by which particular behavioral responses acquire their levels of salience, or prepotency. Cohen, Braver, and colleagues refer to behavioral learning principles to account for how associations are formed between particular pieces of context information and specific outcomes (Braver et al., 1999), linking context information to incentive salience and therefore to behavioral response salience. However, they do not account for the

initial identification and categorization of pieces of information (unless a stochastic process of sampling reward value from among the set of available stimuli is assumed), nor do they argue that associations between behaviors, context information, and outcomes will generalize across situations. This ambiguity aside, as we shall discuss, this model provides critical theoretical traction in our attempt to understand how information processing abnormalities might contribute to the manifestation of thought disorder.

Studies of Information Processing Deficits Related to Formal Thought Disorder

To help us fill in the gap in available theory between mechanisms underlying information processing ability and the mechanisms generating organized, goal-directed speech, we turn to the literature on experimental approaches to studying the pathology of formal thought disorder. Thankfully, this work has been examined capably in a meta-analysis performed by Kerns and Berenbaum (2002), who organize the range of published hypotheses involving specific cognitive impairments associated with formal thought disorder into four general categories.

The first of these categories includes investigations of cognitive mechanisms relatively proximal to speech production (e.g., Barch & Berenbaum, 1996; Goldberg et al., 1998). As alluded to earlier, Kerns and Berenbaum (2002) report only a very minor relationship between impairment in what Levelt would consider the phonological/phonetic system (Indefrey & Levelt, 1999) and ratings of thought disorder. Furthermore, they argue that this tenuous relationship is carried entirely by measures of anomia and word substitution and approximation, deficits likely related to the retrieval of lemmas from the mental lexicon (Indefrey & Levelt, 1999). The vast majority of clinical phenomena related to formal thought disorder (Andreasen & Grove, 1986), however, is left unaccounted for by deficient speech production.

Kerns and Berenbaum's (2002) second category of hypothesized deficit involves increased amount of activation spreading automatically between nodes in semantic networks, resulting in increased priming of nearby semantic associates of a target word, raising the probability that one of these nontarget words will be retrieved and integrated into ongoing speech. This is a relatively intense area of study in schizophrenia research (for a review, see Minzenberg, Ober, & Vinogradov, 2002). Investigators looking

for evidence of abnormal semantic network priming have reported seemingly contradictory findings, with some showing evidence of hyperpriming at tested nodes (suggesting increased amount of activation spreading throughout the network; Spitzer, Brauh, Hermle, & Maier, 1993; Spitzer et al. 1994; Weisbrod, Maier, Harig, Himmelsbach, & Spitzer, 1998) and others showing evidence of hypoprimeing at tested nodes (suggestive of a reduced amount of activation; Besche et al., 1997; Blum & Freides, 1995; Kostova, Passerieux, Laurent, & Hardy-Bayle, 2003). These contradictory data prompted the suggestion that thought disordered patients actually experience an increase in *distance* of automatic activation spread, while maintaining an overall level of activation comparable with controls, effectively yielding an increased number of nodes activated, with none activated to as high of a degree as controls' nodes (Spitzer et al., 1994). Kuperberg, Kreher, and Ditman (2010) provide an informative review of semantic activation studies among schizophrenia patients and frame this activation distance versus activation intensity argument in terms of automatic spread of activation (which is increased in distance among patients) and controlled retrieval of semantic information (a circumstance in which patients show decreased activation in semantic networks).

Along these lines, Kerns and Berenbaum (2002) highlight the contribution of impairment in the controlled retrieval of information from semantic memory, which may itself have an abnormal structure due to the cumulative effects of a chronic inability to encode semantic information. Relevant studies (e.g., Allen, Liddle, & Frith, 1993; Goldberg et al., 1998; Kerns & Berenbaum, 2002) tend to employ fluency tasks, requiring retrieval of information from semantic memory by means such as a controlled implementation of a retrieval strategy (Jacoby, 1991; Ruff, Light, Parker, & Levin, 1997). In agreement with conclusions offered by Minzenberg and colleagues (2002) and Baving and colleagues (2001), who argue that semantic retrieval is most consistently and robustly impaired in schizophrenia patients when a high degree of controlled processing is required, Kerns and Berenbaum (2002) present evidence of a strong, consistent association between this type of semantic processing abnormality and presence of formal thought disorder, especially when evident as poverty of speech.

Evidence of impaired semantic retrieval associated with formal thought disorder suggests a failure in the smooth integration of stored information

with incoming information and response selection. Specifically, information from (long-term) semantic memory is continuously retrieved and integrated into comprehension and online production of verbal behavior. Failure of this fluid integration therefore may recruit more capacity-limited, controlled processing resources, likely involving activation of left inferior prefrontal cortex, to facilitate the otherwise automatic selection of semantic information (Gold & Buckner, 2002; Indefrey & Levelt, 1999). Interestingly, Kerns (2007) has demonstrated in work with schizophrenia patients that impaired performance in an experimental controlled retrieval task correlates specifically with poverty of speech, or what Andreasen and Grove refer to as "negative thought disorder."

This process of semantic memory retrieval and integration itself is very likely to be modulated (see Ragland, Yoon, Minzenberg, & Carter, 2007) by the subject of Kerns and Berenbaum's (2002) fourth category of cognitive deficit contributing to formal thought disorder, namely, impaired executive functioning. As a composite construct, Kerns and Berenbaum demonstrate that executive function abnormality is strongly related to the presence of formal thought disorder. Of course, executive function itself entails a number of critical subsystems (Baddeley, 1986), including a mechanism for processing context information (and effectively inhibiting irrelevant, noncontext information), a mechanism for allocation of attentional capacity serving to maintain information over a delay, and a mechanism for monitoring one's own behavior, including speech.

CONTEXT/SELECTIVE ATTENTION

Consistent with Cohen's and Braver's implementation of the guided activation model, there is considerable evidence that thought-disordered patients suffer from abnormal processing of context information. In fact, Levelt's model of speech production incorporates context information at numerous stages, such as during conceptual preparation (when interpersonal context is considered, for instance). In addition, the process of lexical selection may be influenced by discourse context (Horn & Ward, 2001), which describes the representation of previously uttered verbal information one must hold in mind in order to ensure that subsequent utterances will show adequate structural continuity with and semantic and conceptual relevance to the overarching conversation (Ditman & Kuperberg, 2010).

Numerous investigators have examined schizophrenia patients' capacity to use discourse context to guide selection of verbal behaviors. For instance, a significant body of work (Benjamin & Watt, 1969; Cohen & Servan-Schreiber, 1992; Kuperberg, McGuire, & David, 1998; Sitnikova, Salisbury, Kuperberg, & Holcomb, 2002) involves using various lexical disambiguation tasks, which require the subject to use context information from preceding clauses to determine the relevant meaning of a homograph, or a word with multiple possible definitions. These and other investigators have generally concluded that patients with psychotic disorders fail to demonstrate sensitivity to the biasing influence of preceding context information. Notably, this conclusion complements Chapman's and Chapman's (1973) argument that patients fail to demonstrate sensitivity to discourse context only when it suggests a homograph's nondominant meaning. They characterized this deficit as "excessive yielding to normal biases," or a tendency to utilize dominant meanings—what Ditman and Kuperberg (2010) refer to as an overreliance on lexico-semantic associations, rather than discourse context. For instance, when one patient was asked to interpret the proverb "One swallow does not make a summer," he responded, "When you swallow something, it could be all right, but the next minute you could be coughing, and dreariness and all kind of miserable things coming out of your throat" (quoted in Harrow & Quinlan, 1985, p. 436). The patient clearly demonstrated a bias toward the more dominant meaning of the word "swallow," despite the fact that the context of the question implied the nondominant meaning.

Of course, excessive yielding to normal biases is the logical complement of Cohen's and Braver's cognitive control mechanism, which is defined by its ability to overcome these normal biases. Accordingly, Cohen, Braver, and colleagues argue (Braver et al., 2001) that an individual's representation of discourse context, as well as his or her goals for the interaction (e.g., make a particular point, communicate in a certain manner) comprise context information, guiding the ongoing implementation of related semantic concepts. Failure to encode, update, and/or maintain this context information therefore leads to a failure to utilize discourse context to constrain and select subsequent verbal output, appearing to the observer as a relative lack of association between units of language output.

Moreover, if failure to encode, maintain, or implement context information is in fact a mechanism

underlying formal thought disorder, it may explain a long-held piece of clinical wisdom: Disordered speech is more likely to be elicited by abstract, ambiguous, open-ended stimuli (such as the general question posed to the quoted subjects at the beginning of this chapter, or even Rorschach inkblots; Johnston & Holzman, 1979) than by specific, closed-ended prompts. In other words, the fewer structural demands and intermediate goal states provided explicitly, the more difficult it is to practice cognitive control. Under these circumstances, not only is specific context information either never encoded or lost from active maintenance, but the context-processing module loses the concomitant ability to inhibit the activation of competing pieces of information, exposing the system to increased memory retrieval interference (Anderson & Spellman, 1995) and subsequent loss of goal orientation in produced speech. Among nonpsychiatric control subjects, Kerns and colleagues (2004) demonstrated that increased encoding and maintenance of context information, associated with increased prefrontal activation, predicted subsequent use of context-appropriate responses.

CAPACITY ALLOCATION

Given this continued focus on cognitive control as critical to information processing abnormalities related to formal thought disorder, it is important to discuss the allocation of working memory capacity, a process shown to involve activation of dorsolateral prefrontal cortex as well as more modality-specific regions of posterior cortex (e.g., Garavan, Ross, Li, & Stein, 2000), as well as capacity availability itself, which appears to be reflected in the activity of dorsolateral (Callicott et al., 1999) and ventrolateral prefrontal cortex (Rypma, Berger, & D'Esposito, 2002). Numerous studies (e.g., Docherty & Gordinier, 1999; Harvey & Pedley, 1989; Nuechterlein & Dawson, 1984) have found correlational evidence of a relationship between working memory capacity and aspects of formal thought disorder. Attempting to clarify the direction of this relationship, Barch and Berenbaum (1994) report that, among non-ill subjects, reduction in overall processing capacity (achieved through a dual-task manipulation) is associated with decreases in verbosity and syntactic complexity, verbal phenomena included in formal thought disorder. Kerns (2007) followed this work up more recently in specifying that disorganized speech—separate from impoverished speech—is most strongly associated with experimental measures of reduced working

memory capacity among schizophrenia patients. Melinder and Barch (2003) extended this approach to include psychotic patients, showing that they too show increased negative thought disorder with decreasing availability of working memory capacity. These results are particularly noteworthy because the investigators were able to demonstrate, at least under highly controlled circumstances employing a model system, that reduced processing capacity can actually cause speech to become disordered. Whether this mechanism accounts for the observed correlation between reduced processing capacity and thought disorder among patients remains to be determined. Nevertheless, this represents one instance out of many in which schizophrenia research has suggested that working memory capacity may act as a bottleneck, limiting the production and/or implementation of abstract ideas (e.g., Glahn, Cannon, Gur, Ragland, & Gur, 2000; Silver, Feldman, Bilker, & Gur, 2003).

SELF-MONITORING

Levett (1989) argues that effective communication requires the speaker to monitor his or her own speech and to self-correct any erroneous utterances. Interestingly, schizophrenia patients show an impairment in the ability to self-correct erroneous behaviors (e.g., Malenka, Angel, Hampton, & Berger, 1982), which may result from patients' failure to notice errors spontaneously or their failure to recruit additional cognitive control resources to recover from errors and cope with increased task difficulty (van Veen & Carter, 2006). Alternatively, uncertainty regarding the source of erroneous speech may also explain patients' failure to self-correct. Several investigators have shown that patients have difficulty determining whether speech was self-generated, especially under high distortion (e.g., Kumari et al., 2010). In at least one case, this failure was significantly associated with a greater number of verbal derailments (i.e., switching tangentially between topics of discussion; Barch & Berenbaum, 1996).

INTEGRATION OF COGNITIVE DEFICITS CONTRIBUTING TO FORMAL THOUGHT DISORDER

Taken together, these findings suggest a model in which deficits in semantic memory retrieval and executive functions contribute to the expression of formal thought disorder. A study published by Titone and colleagues (2000) provides an example

of this type of interaction. The authors reported the results of a semantic priming study in which particular meanings of otherwise semantically ambiguous words were biased either moderately or strongly by the context of a preceding sentence. The presentation of potentially biasing semantic information was carried out in a manner to increase the likelihood that more controlled retrieval of semantic information would be utilized. Schizophrenia patients showed a pattern of priming identical to controls in the strong context bias condition but exhibited a greater degree of priming in the moderate context bias condition (i.e., patients' showed priming effects for both relatively dominant and relatively subordinate meanings, while controls showed priming facilitation only for subordinate meanings). The authors point out that retrieval of a particular meaning of a word requires not only activation of the word within a semantic network but also inhibition of nearby, less relevant meanings. Patients were able to perform this selection process normally when strong context bias was present, but when this influence was more subtle, patients' degraded retrieval-related inhibitory mechanism failed to filter out alternate meanings, creating interference with the most immediately relevant meaning. Similarly, Ditman and Kuperberg (2010) extended the approach to examine discourse context across sentences, showing that patients successfully maintain and utilize very strong, highly restrictive context; however, patients again failed to use less restrictive context information in subsequent sentences. Therefore, to the extent that the Ditman and Kuperberg study indeed engaged controlled processing mechanisms and therefore did not rely entirely on the automatic spread of activation in a semantic network, the results support the hypothesis that disordered speech results from disrupted executive-assisted semantic memory retrieval mechanisms involving both abnormal activation-based retrieval of information from semantic memory and impaired inhibition of irrelevant, non-context information.

Ex Cogito, Dementia

As the foregoing discussion illustrates, cognitive characterization of thought disorder has many merits, not the least of which is the ability to predict patient performance data in a variety of experimental cognitive tasks. In addition, these models converge with descriptive analyses of the experience of thought disorder in patients with psychotic disorders. Nevertheless, demonstrating that a particular

neurocognitive impairment *could* account for a particular behavioral abnormality does not necessarily demonstrate that the impairment *does* cause the abnormal behavior to occur.

Indeed, more than four decades of intensive neuroscientific investigation have failed to identify conclusively a single defining lesion in patients with schizophrenia or other forms of psychosis. Rather, these syndromes manifest with deficits to many neural systems across several levels of analysis. In light of this complexity, we apply an analytic framework that has become the dominant paradigm in psychopathology research—that is, the endophenotype approach (Gottesman & Gould, 2003)—to theoretical accounts of information processing impairment, including thought disorder. The basic premise of the endophenotype approach is that a given clinical syndrome such as schizophrenia is composed of multiple neurocognitive trait deficits, each of which may be determined by at least partially independent mechanisms. A major consequence of this model is that a certain trait deficit may be necessary but not sufficient for the phenotypic manifestation of a syndrome; thus, the trait deficit will be shared by individuals with a vulnerability to the syndrome (e.g., patients' biological relatives), regardless of whether they manifest the syndrome phenotypically. Other deficits may be specific to individuals who manifest the syndrome phenotypically; these latter deficits may thus potentiate the expression of a symptom in those who carry vulnerability (i.e., those who have deficits in other neurocognitive domains that are necessary but not sufficient for overt disease expression). To develop this framework further in the context of a discussion of thought disorder, it will first be useful to explicate a number of facts about the genetic epidemiology and clinical neuroscience of schizophrenia (also see Green & Dunbar, Chapter 7).

The Genetic Epidemiology of Schizophrenia

Although we are aware of only very limited work on the heritability of formal thought disorder itself (Levy et al., 2010), a great deal of evidence is available demonstrating that genetic factors contribute substantially to the development of schizophrenia, accounting for about 80% of liability to developing the disorder. The transmission pattern, however, is complex, involving at least several different genes as well as environmental factors (Cannon, Kaprio, Lonnqvist, Huttunen, & Koskenvuo, 1998; Tsuang & Faraone, 1999; Tsuang, Stone, & Faraone, 1999).

One consequence of the complexity of the inheritance pattern in schizophrenia is that an individual may carry some degree of genetic predisposition to the illness without expressing it phenotypically—or at least without expressing it to a degree severe enough to meet diagnostic criteria. Stated differently, only a subset of genetically vulnerable individuals actually develops a psychotic disorder. For many with such a genetic predisposition, an environmental contribution (to which genetically predisposed individuals might be differentially sensitive) to development of a psychotic disorder is also required. Among the environmental factors that may be involved, prenatal and perinatal complications, particularly those associated with fetal hypoxia, or oxygen deprivation, are robustly associated with an increased risk for schizophrenia. Complications associated with fetal hypoxia are also of interest because fetal oxygen deprivation represents a plausible mechanism for explaining much of the structural pathology of the brain detected in neuroimaging studies of adult schizophrenia patients (Cannon, 1997).

Applying the conclusion that such genetic and environmental influences aggregate together (additively or interactively) to determine an individual's risk for expressing a psychotic disorder to the study of neurocognitive traits helps demonstrate which of such traits are likely necessary, but not sufficient, for the expression of a psychosis phenotype (or to the expression of any phenotype, including specific symptoms, for example). Specifically, deficits related entirely to the genetic diathesis for developing the given phenotype may be necessary but clearly are not sufficient for the manifestation of that phenotype. This endophenotype should be present in any individual carrying the genetic vulnerability. Consequently, if one member of a set of monozygotic twins (who, by definition have identical genomes) displays a vulnerability-specific trait, the other must as well. In addition, any trait not shared by both monozygotic twins must result to some degree from the influence of unique environmental events (in particular, those related causally to the disorder or those that reflect consequences of having the disorder or being treated for it).

Neural System Abnormalities in Schizophrenia

While neither the specific neurobiological processes associated with the expression of formal thought disorder nor those associated with psychosis in general have been definitively isolated,

disturbances in prefrontal and temporo-limbic systems and their interconnections are likely to play critical roles in both (Cohen & Servan-Schreiber, 1992; Grace, Moore, & O'Donnell, 1998; Gray et al., 1991). The prefrontal cortex—the region of the brain most often associated with formal thought disorder (Goghari, Sponheim, & MacDonald, 2010)—is believed to support higher order cognitive processes such as working memory, the strategic allocation of attention, reasoning, planning, and other forms of abstract thought (Goldman-Rakic, 1995; Kane & Engle, 2002; Miller & Cohen, 2001; also see Morrison & Knowlton, Chapter 6). Medial temporal lobe structures (i.e., hippocampus, amygdala) and adjacent temporal cortex are involved in learning and recall of episodic and other relational information, emotion (especially the amygdala), and certain aspects of language processing (Squire & Zola, 1996).

Neuropsychological studies have shown that, against a background of generalized information processing impairment, schizophrenia patients show profound deficits in the areas of both long-term and working memory (Cannon et al., 2000; Saykin et al., 1994). These deficits appear not to be merely secondary effects of impaired attention, disease chronicity, or medication exposure (Cirillo & Seidman, 2003). Such findings have been corroborated by evidence of abnormal physiologic activity (i.e., altered blood flow) in prefrontal and temporal lobe regions in patients with schizophrenia during performance of tests assessing these same domains of functioning (Berman, Torrey, Daniel, & Weinberger, 1992; Callicott et al., 1998; Heckers et al., 1998; Yurgelun-Todd et al., 1996). At the structural-anatomical level, schizophrenia patients show a variety of volumetric changes throughout the brain, including reduced cortical, hippocampal, and thalamic volumes (Shenton, Dickey, Frumin, & McCarley, 2001). Recent neuroimaging work demonstrates a greater degree of reduction in frontal and temporal cortical volumes compared with posterior cortical volumes (Sun, Stuart et al., 2009; Sun, van Erp et al., 2009). Additionally, the white matter tracts supporting communication between these critical frontal areas and other modality-specific regions are affected (Karlgodt et al., 2008; Rosenberger et al., 2008).

PREFRONTAL CORTEX AND WORKING MEMORY DEFICITS

Several lines of evidence suggest that working memory deficits and associated abnormalities in

prefrontal cortical structure and function are reflective of an inherited diathesis to schizophrenia. In a Finnish twin sample, Cannon and colleagues found that impaired performance on tests of spatial working memory capacity and structural abnormalities in polar and dorsolateral prefrontal regions varied in a dose-dependent fashion with degree of genetic loading for schizophrenia (Cannon et al., 2000, 2002; Glahn et al., 2002). Interestingly, global and dorsolateral prefrontal volumetric deficits have been found to correlate with performance deficits on tests sensitive to diverse working memory processes (Maher, Manschreck, Woods, Yurgelun-Todd, & Tsuang, 1995; Seidman et al., 1994).

Initial functional neuroimaging studies examining the neural correlates of patients' working memory deficits provided what appeared to be highly discrepant findings, with some reporting prefrontal hypoactivation among patients and others reporting relative hyperactivation (see Karlsgodt et al., 2009, for a review). However, careful consideration of these results demonstrated that at least three factors may play a critical role in modulating prefrontal activity. Among these are individual and group differences in functional and structural connectivity between prefrontal regions and other neural structures recruited during working memory performance (Glahn et al., 2005; Karlsgodt et al., 2008), reinforcing the notion that no locus of brain activity can be characterized entirely without consideration of its interconnectivity with other regions.

Behavioral performance achieved during task performance also has received increased scrutiny (Callicott et al., 2003; Manoach, 2003), with at least one demonstration that behavioral performance may statistically moderate prefrontal activation, placing the relative hypo- versus hyperactivation debate squarely within the context of task performance (Van Snellenberg, Torres, & Thornton, 2006). Finally, when behavioral performance is controlled, the efficiency of neural activation—the relative change in neural activity per unit increase in behavioral performance—has drawn attention. In fact, a robust group-wise difference in the function associating behavior and neural activation in a given brain region (Karlsgodt et al., 2007) offers a physiological explanation for decreased activation among patients failing to maintain information in working memory and for increased activation among patients performing working memory tasks successfully. Figure 34.3 illustrates this phenomenon, whereby patients' accurate performance on a spatial

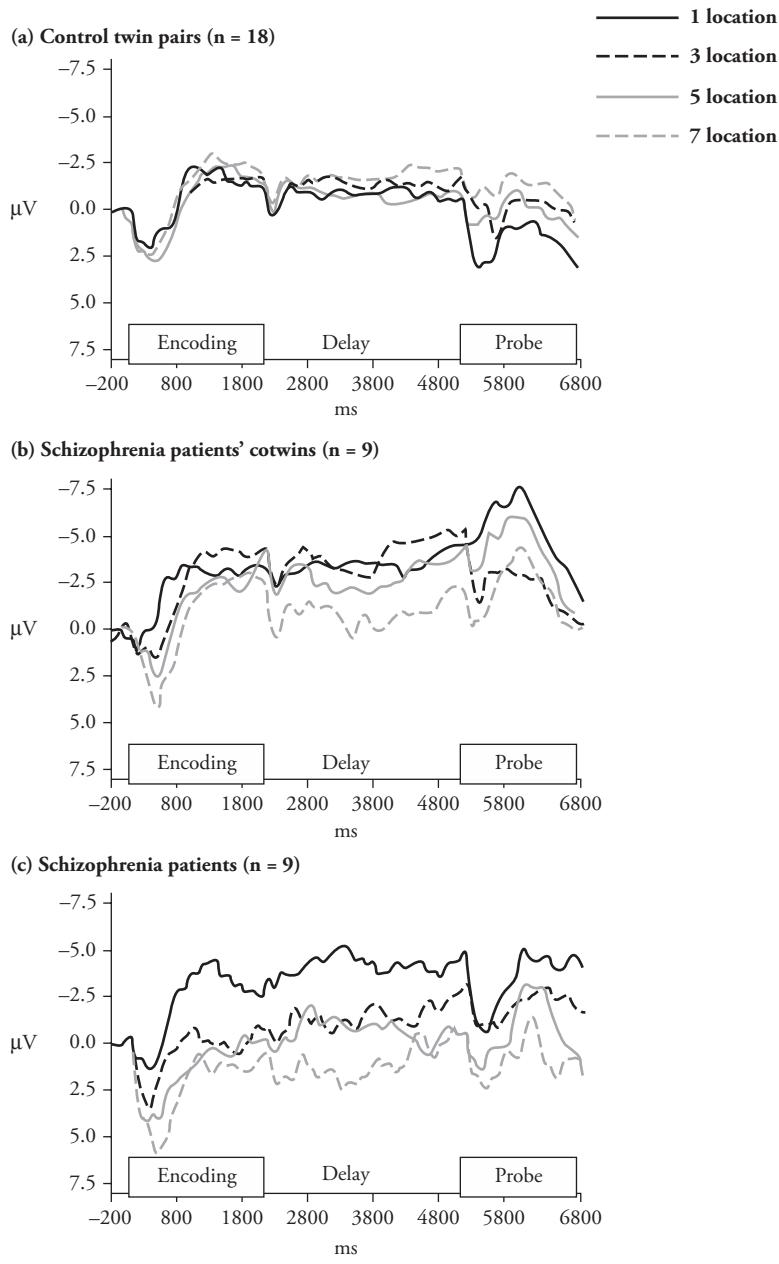


Fig. 34.3 Event-related potential (ERP) waveforms at a centroparietal electrode location, elicited by memory set stimulus onset, displayed at each memory load level for controls (*a*), patients' co-twins (*b*), and schizophrenia patients (*c*). Waveforms were low-pass filtered at 16 Hz (24-dB/oct roll-off) for display purposes only (from Bachman et al., 2009).

working memory task engages a much greater range of slow-wave event-related potential activity than controls show, with patients' non-ill co-twins intermediate (Bachman et al., 2009).

However, the nature of the pathological mechanism underlying these abnormal brain-behavior relationships is not obvious. Rather than a loss of neurons or interneurons, it has been suggested

that gross gray matter volume decrements reflect a reduction of interneuronal neuropil—the space between neural cells, consisting largely of dendrites and synapses—in the prefrontal region in patients with schizophrenia, resulting in impaired working memory functioning due to hypoactive dopaminergic modulation of pyramidal cell activity (Goldman-Rakic & Selemon, 1997). Rather than

subcortical dopaminergic dysregulation, in this case dopamine would be acting within the cortex (although affecting a distinct set of receptors). This prediction has been supported by a positron emission tomography (PET) investigation that found significantly decreased dopamine receptor binding in the prefrontal cortex of schizophrenia patients (Okubo et al., 1997). Notably, dopamine receptor reduction predicted certain types of symptoms, as well as working memory impairment (but also see Abi-Dargham et al., 2002). It is also of interest in this context that treatment with medication modulating cortical dopamine levels is associated with normalization of blood flow in the prefrontal cortex and increased behavioral accuracy during performance of a working memory test (Honey & Watt, 1999).

Another, not mutually exclusive mechanism potentially explaining a significant degree of dysfunction involves the parvalbumin-expressing subpopulation of GABA neurons in the dorsolateral aspect of the prefrontal cortex, which appear to play a privileged role in regulating the timing of oscillatory activity within localized cortical circuits, and also reach mature levels of functioning along a trajectory that appears to mirror increased incidence of schizophrenia and other psychotic illnesses (Lewis & González-Burgos, 2008).

Given that abnormalities of working memory and prefrontal structure and function are associated with genetic liability to schizophrenia, it should be possible to identify specific genes that underlie these disturbances, especially in light of accumulating evidence of physiological abnormality. Weinberger and colleagues have reported on a number of associations between variations at particular genetic loci and prefrontal activation during working memory performance (Tan, Callicott, & Weinberger, 2009). Among these, the MET/VAL polymorphism of the COMT gene (located on chromosome 22) may modulate the efficiency of prefrontal activation, with VAL alleles promoting more rapid breakdown of synaptic dopamine leading to prefrontal hypo-function in patients with schizophrenia (Egan et al., 2001). However, this line of work has also demonstrated that even attempts to identify "intermediate phenotypes" that map particular genes to the physiological networks supporting behavior inevitably meet with considerable complexity, including epistatic interactions between target genes and expression of functionally related molecules (e.g., COMT and GAD1 co-expression; Straub et al., 2007).

We have been interested in another potential susceptibility locus that may affect prefrontal function in schizophrenia, this one on chromosome 1. Cannon et al. (2005) found that inherited variations in the DISC1 gene are associated with schizophrenia diagnosis, reduced prefrontal and medial temporal lobe gray matter, and impaired sociability and memory functioning. Furthermore, using an inducible transgenic animal model, Li and colleagues (2007) demonstrated that interfering with DISC1 during development results in reductions in dendritic complexity and synaptic density as well as impairments in sociability and working memory, paralleling the findings in patients. Taken together, these findings support the view that inherited disruptions in neural connectivity affecting gray matter underlie important dimensions of schizophrenia, including disruptions in working memory.

Together, these findings strongly implicate genetic factors as playing a role in the abnormalities of prefrontal cortex and working memory in schizophrenia. Because deficits on tests sensitive to working memory have also been observed in children at elevated genetic risk (Cosway et al., 2000), it is tempting to conclude that disturbances in the prefrontal cortex in schizophrenia are reflective of an inherited vulnerability to the disorder that is present from early in life. Nevertheless, patients with schizophrenia have been found to show even greater disturbances in dorsolateral prefrontal cortex function and structure than their non-ill monozygotic co-twins (Cannon et al., 2002). Thus, while genetic factors may cause patients and some of their first-degree relatives to share a certain degree of compromise in prefrontal cortical systems, nongenetic disease-specific influences cause the dorsolateral prefrontal cortex to be further deviant in the patients.

TEMPORAL LOBE AND EPISODIC MEMORY DEFICITS

Several microscopic abnormalities of the hippocampus have been documented in schizophrenia, including alterations in neuronal density (Falkai & Bogerts, 1986; Huttenlocher, 1979; Jeste & Lohr, 1989; Zaidel, Esiri, & Harrison, 1997), size (Arnold, 2000; Benes, Sorensen, & Bird, 1991), and orientation (Conrad, Abebe, Austin, Forsythe, & Scheibel, 1991; Conrad & Scheibel, 1987; Kovelman & Scheibel, 1984). These hippocampal volume decrements appear to be present at disease onset (Bilder et al., 1995; Velakoulis et al., 1999) and also appear to be present to some degree in healthy biological

relatives of schizophrenia patients, suggesting hippocampal volume is related to the genetic diathesis for developing schizophrenia (Lawrie et al., 1999; Seidman et al., 1999, 2002). Postmortem and magnetic resonance imaging (MRI) studies of schizophrenia patients, however, have reported positive correlations between hippocampal volume and age at onset (Bogerts et al., 1990; Dauphinais et al., 1990; Stefanis et al., 1999; Van Erp et al., 2002), suggesting a relationship between hippocampal volume and the disease process, complicating any simple interpretation. One possible explanation is that only very subtle structural abnormalities (Wood et al., 2010) coincide with significant metabolic dysregulation (Schobel et al., 2009) of critical hippocampal subfields prior to psychosis onset. The initial episode of psychosis may then alter the structure of the hippocampus, leaving it vulnerable to factors that aggregate with illness chronicity (Velakoulis et al., 2006).

From a neurocognitive perspective, impaired declarative memory processes that depend on the integrity of the hippocampus (Faraone et al., 2000) have been examined extensively among adult schizophrenia patients. In particular, patients show greatest mnemonic deficits when learning depends upon relational encoding between a target stimulus and contextual information (van Erp et al., 2008).

Similar behavioral impairments in long-term memory have been reported in both high-risk adolescents (Byrne, Hodges, Grant, Owens, & Johnstone, 1999; Seidman et al., 2010) and nonpsychotic relatives of schizophrenia patients (Cannon et al., 1994), suggesting that they derive in part from an inherited genotype. However, because long-term memory deficits are specifically more pronounced in patients compared with their own healthy monozygotic co-twins, nongenetic, disease-specific factors must also be involved (Cannon et al., 2000; van Erp et al., 2008). Importantly, two studies have shown a significant relationship between deficits in verbal declarative memory and smaller hippocampal volumes in relatives of schizophrenia patients (O'Driscoll et al., 2001; Seidman et al., 2002).

Given the putative importance of the hippocampus to verbal and executive function, and therefore its possible role in producing disordered speech, it is of interest to revisit the issue of genetic versus environmental contributions to hippocampal integrity. Compared to other parts of the brain, the hippocampus is acutely vulnerable to hypoxic-ischemic damage (Vargha-Khadem et al., 1997; Zola & Squire, 2001), that is, insult temporarily depriving neural cells of

oxygen. In monozygotic twins discordant for schizophrenia, relatively reduced hippocampal volume in the ill twin was significantly related to the presence of labor-delivery complications and to prolonged labor, both risk factors associated with fetal oxygen deprivation (McNeil, Cantor-Graae, & Weinberger, 2000). Van Erp and colleagues found, in a Helsinki birth cohort, that schizophrenia patients who experienced fetal hypoxia have smaller hippocampal volumes than in those who did not, a difference not noted within unaffected siblings and healthy comparison subjects (Van Erp et al., 2002). At the same time, hippocampal volume differences occurred in a stepwise fashion with increase in genetic vulnerability for developing schizophrenia (consistent with the findings of Seidman et al., 2002), suggesting that, in patients with schizophrenia spectrum disorders, hippocampal volume is influenced in part by schizophrenia susceptibility genes and an interaction of these genes with experience of fetal hypoxia. Together, these findings indicate that while hippocampal volume in healthy subjects is under substantial genetic control, hippocampal volume in schizophrenia patients and their relatives appears to be influenced to a greater extent by unique and shared environmental factors (Van Erp et al., 2002).

Integrating Cognitive Models and Endophenotypes

A number of useful links exist between the cognitive models of disrupted information processing in schizophrenia patients reviewed in the first part of this chapter and the research on neurocognitive endophenotypes in schizophrenia just summarized. At the cognitive level of analysis, there appear to be two mechanisms necessary for the expression of formal thought disorder: an executive, online processing system, responsible for the encoding, maintenance, and updating of goal-related information (context information in the Guided Activation model), and an integrated system involving the consolidation and storage of semantic and episodic information (Kerns, 2007).

In terms of the endophenotype framework developed earlier, individuals at elevated genetic risk, but not expressing the schizophrenia phenotype, show mildly impaired functioning of executive systems and related working memory and attention components. These executive processing deficits therefore appear to be associated with the diathesis, which is necessary but not sufficient for the development of thought disorder. Beyond this diathesis, the

abnormal interaction of executive and semantic memory systems—likely in service of controlled retrieval of information, and its integration into coherent, logically consistent discourse—is associated with a psychosis-specific factor, itself related to both genetic vulnerability and exposure to environmental risk factors. Individuals with schizophrenia and their unaffected co-twins show a qualitatively similar pattern of prefrontal structural and functional abnormality, somewhat greater in severity in the patients. Patients and their relatives additionally show temporal lobe abnormalities; however, the degree of difference in temporal lobe abnormality between schizophrenia patients and genetically vulnerable individuals is significantly larger than the corresponding difference in prefrontal abnormality.

Taken together, these results suggest that mild impairment in prefrontal cortex and associated degradation of online cognitive processing systems (i.e., executive functions, including working memory and selective attention), comprise a necessary but not sufficient (i.e., contributing) cause of thought disorder, which itself derives from a genetic diathesis to developing a psychotic disorder such as schizophrenia. An additional factor, related etiologically to exposure to an environmental insult interacting with genetic predisposition, and also necessary but not sufficient for the expression of schizophrenia, involves disrupted interaction between an executive, online processing system and a semantic memory storage and selection system, loosely mapping onto schizophrenia patients' prefrontal and temporal lobe abnormalities, respectively. That is, abnormalities in both the prefrontal/executive-related circuitry and in the temporal lobe circuitry (i.e., medial temporal lobe for episodic memory and nearby middle temporal gyrus for semantic memory; Kircher et al., 2001) may be required to account for the full range of thought disorder observed in patients with schizophrenia, while only the former may be required to account for the subtler thought disturbances seen in genetically vulnerable individuals who do not manifest the full schizophrenia syndrome phenotypically. Of course, it is also possible that severity of phenotypic thought disorder scales with severity of compromise of both components of the system, rather than to their conjunction *per se*. Further work is needed to segregate these two possibilities.

Conclusions and Future Directions

In summary, research related to the cognitive, genetic, and neural pathologies of thought disorder

in general, and schizophrenia specifically, has necessarily taken on a complex, interactive structure. As we have seen, cognitive models designed to predict particular behavioral outcomes can in fact help researchers understand the functional correlates of anatomical abnormalities measured between genetically defined risk groups. Similar permutations involving these and numerous other levels of analysis equip us with heuristics that guide our struggle to unravel the complexities of neuropsychiatric phenomena such as formal thought disorder. We have attempted to present such a heuristic framework, based on links we have observed between bodies of research into the pathology of thought disorder, some of these links crossing between levels of analysis, ideally helping us to map genetic, neurological, and cognitive systems onto each other.

Along the way to accomplishing this integrative goal, a great deal more work needs to be done. Ideally, the parsing of formal thought disorder into necessary and sufficient functional components will be complemented by further study of the physiological and genetic variations associated with the production of abnormal speech. This line of work will likely be facilitated by cognitive neuroscience's growing ability to study the activity of particular brain mechanisms during the production of speech, overcoming previously prohibitive practical obstacles caused by movement artifacts detrimental to work utilizing common neuroscientific research modalities (e.g., Costa, Strijkers, Martin, & Thierry, 2009). Prior to these methodological advances, only speech production studies employing *covert* vocalization were practical; however, these investigations typically fall short of describing compellingly aspects of formal thought disorder itself, to a phenomenon measured entirely in terms of *overt* speech production.

Additional progress in the study of thought disorder involves application of paradigms from the emerging field of social cognitive neuroscience (e.g., Adolphs, 2003; Wood, 2003) to the study of interpersonal deficits in schizophrenia (e.g., Penn, Ritchie, Francis, Combs, & Martin, 2002; Pinkham, Penn, Perkins, & Lieberman, 2003), including the distinctly interpersonal task of verbal communication (Grossman & Harrow, 1996; Racenstein, Penn, Harrow, & Schleser, 1999). For instance, the study of communication deviance (including aspects of formal thought disorder) within the families of patients with psychotic disorder diagnoses, or patients thought to be at high risk for developing a psychotic disorder, has been an area of active

research for some time (e.g., Docherty, 1995; Sass, Gunderson, Singer, & Wynne, 1984; Wahlberg et al., 2000). Applying this established framework to the examination of neuronal correlates of receptive and productive aspects of intrafamily communication—potentially distinct from communication with nonfamily members due to the role of factors such as increased interpersonal familiarity and less predictable affective modulation of cognitive processes involved in communication—offers a novel perspective with the potential to reinvigorate this important line of thought disorder research.

Another area of thought disorder research deserving of continued attention involves the study of formal thought disorder in populations other than those currently meeting diagnostic criteria for a major mental illness. Although modern antipsychotic medications appear to be relatively effective at helping psychotic patients organize their speech (e.g., Wirshing et al., 1999), significant levels of thought disorder often appear noticeable in groups of patients who would not typically be treated with therapeutic doses of such medications (Andreasen & Grove, 1986).

For instance, examination of a large sample of individuals judged to be at significantly elevated risk for developing a psychotic disorder (in part because they were displaying some psychotic symptoms, only at a level of intensity and/or frequency below diagnostic threshold) found a significant level of formal thought disorder (Woods et al., 2009). Furthermore, communication disturbance in this population has been linked to attentional impairment (Cannon et al., 2002) and to elevated, inefficient activation of prefrontal language areas during a discourse monitoring task (Sabb et al., 2010). These and similar findings raise interesting questions regarding the potential utility of formal thought disorder as a prodromal indicator of psychosis, as well as the potential benefits of symptom-based treatment outside the context of a major psychiatric diagnosis.

Acknowledgments

Preparation of this manuscript was supported by grants MH52857, MH65079, and MH66286 from the National Institute of Mental Health (NIMH) and by gifts to the UCLA Foundation from Garen and Shari Staglin and the International Mental Health Research Organization (IMHRO).

References

- Abi-Dargham, A., Mawlawi, O., Lombardo, I., Gil, R., Martinez, D., Huang, Y., et al. (2002). Prefrontal dopamine D1 receptors and working memory in schizophrenia. *Journal of Neuroscience*, 22(9), 3708–3719.
- Abrahams, B. S., & Geschwind, D. H. (2010). Connecting genes to brain in the autism spectrum disorders. *Archives of Neurology*, 67(4), 395–399.
- Adolphs, R. (2003). Cognitive neuroscience of human social behavior. *Nature Reviews Neuroscience*, 4(3), 165–178.
- Allen, H. A., Liddle, P. F., & Frith, C. D. (1993). Negative features, retrieval processes and verbal fluency in schizophrenia. *British Journal of Psychiatry*, 163, 769–775.
- American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.
- Anderson, M. C., & Spellman, B. A. (1995). On the status of inhibitory mechanisms in cognition: Memory retrieval as a model case. *Psychology Review*, 102(1), 68–100.
- Andreasen, N. C. (1979). Thought, language, and communication disorders. I. Clinical assessment, definition of terms, and evaluation of their reliability. *Archives of General Psychiatry*, 36(12), 1315–1321.
- Andreasen, N. C. (1982). Should the term “thought disorder” be revised? *Comprehensive Psychiatry*, 23(4), 291–299.
- Andreasen, N. C. (1986). Scale for the assessment of thought, language, and communication (TLC). *Schizophrenia Bulletin*, 12(3), 473–482.
- Andreasen, N. C., & Grove, W. M. (1986). Thought, language, and communication in schizophrenia: Diagnosis and prognosis. *Schizophrenia Bulletin*, 12(3), 348–359.
- Andreasen, N. C., Nopoulos, P., O’Leary, D. S., Miller, D. D., Wassink, T., & Flaum, M. (1999). Defining the phenotype of schizophrenia: Cognitive dysmetria and its neural mechanisms. *Biological Psychiatry*, 46(7), 908–920.
- Arnold, S. E. (2000). Cellular and molecular neuropathology of the parahippocampal region in schizophrenia. *Annals of the New York Academy of Science*, 911, 275–292.
- Bachman, P., Kim, J., Yee, C. M., Therman, S., Manninen, M., Lönnqvist, J., et al. (2009). Efficiency of working memory encoding in twins discordant for schizophrenia. *Psychiatry Research*, 174(2), 97–104.
- Baddeley, A. D. (1986). *Working memory*. New York: Oxford University Press.
- Barch, D., & Berenbaum, H. (1994). The relationship between information processing and language production. *Journal of Abnormal Psychology*, 103(2), 241–250.
- Barch, D. M., & Berenbaum, H. (1996). Language production and thought disorder in schizophrenia. *Journal of Abnormal Psychology*, 105(1), 81–88.
- Barch, D. M., Braver, T. S., Nystrom, L. E., Forman, S. D., Noll, D. C., & Cohen, J. D. (1997). Dissociating working memory from task difficulty in human prefrontal cortex. *Neuropsychologia*, 35(10), 1373–1380.
- Baving, L., Wagner, M., Cohen, R., & Rockstroh, B. (2001). Increased semantic and repetition priming in schizophrenic patients. *Journal of Abnormal Psychology*, 110(1), 67–75.
- Benes, F. M., Sorensen, I., & Bird, E. D. (1991). Reduced neuronal size in posterior hippocampus of schizophrenic patients. *Schizophrenia Bulletin*, 17(4), 597–608.
- Benjamin, T. B., & Watt, N. F. (1969). Psychopathology and semantic interpretation of ambiguous words. *Journal of Abnormal Psychology*, 70, 67–714.
- Berenbaum, H., & Barch, D. (1995). The categorization of thought disorder. *Journal of Psycholinguistic Research*, 24(5), 349–376.

- Berman, K. F., Torrey, E. F., Daniel, D. G., & Weinberger, D. R. (1992). Regional cerebral blood flow in monozygotic twins discordant and concordant for schizophrenia. *Archives of General Psychiatry*, 49(12), 927–934.
- Besche, C., Passerieux, C., Segui, J., Sarfati, Y., Laurent, J. P., & Hardy-Bayle, M. C. (1997). Syntactic and semantic processing in schizophrenic patients evaluated by lexical-decision tasks. *Neuropsychology*, 11(4), 498–505.
- Bilder, R. M., Bogerts, B., Ashtari, M., Wu, H., Alvir, J. M., Jody, D., Reiter, G., Bell, L., & Lieberman, J. A. (1995). Anterior hippocampal volume reductions predict frontal lobe dysfunction in first episode schizophrenia. *Schizophrenia Research*, 17(1), 47–58.
- Bleuler, E. (1911/1950). *Dementia praecox, or, the group of schizophrenias*. (J. Zinkin, Trans.). New York: International Universities Press.
- Blum, N. A., & Freides, D. (1995). Investigating thought disorder in schizophrenia with the lexical decision task. *Schizophrenia Research*, 16(3), 217–224.
- Bogerts, B., Ashtari, M., Degreef, G., Alvir, J. M., Bilder, R. M., & Lieberman, J. A. (1990). Reduced temporal limbic structure volumes on magnetic resonance images in first episode schizophrenia. *Psychiatry Research*, 35(1), 1–13.
- Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, 108(3), 624–652.
- Braver, T. S., Barch, D. M., & Cohen, J. D. (1999). Cognition and control in schizophrenia: A computational model of dopamine and prefrontal function. *Biological Psychiatry*, 46(3), 312–328.
- Braver, T. S., Barch, D. M., Keys, B. A., Carter, C. S., Cohen, J. D., Kaye, J. A., et al. (2001). Context processing in older adults: Evidence for a theory relating cognitive control to neurobiology in healthy aging. *Journal of Experimental Psychology: General*, 130(4), 746–763.
- Braver, T. S., Cohen, J. D., Nystrom, L. E., Jonides, J., Smith, E. E., & Noll, D. C. (1997). A parametric study of prefrontal cortex involvement in human working memory. *Neuroimage*, 5(1), 49–62.
- Braver, T. S., Reynolds, J. R., & Donaldson, D. I. (2003). Neural mechanisms of transient and sustained cognitive control during task switching. *Neuron*, 39(4), 713–726.
- Byrne, M., Hodges, A., Grant, E., Owens, D. C., & Johnstone, E. C. (1999). Neuropsychological assessment of young people at high genetic risk for developing schizophrenia compared with controls: Preliminary findings of the Edinburgh High Risk Study (EHRs). *Psychological Medicine*, 29(5), 1161–1173.
- Callicott, J. H., Mattay, V. S., Bertolino, A., Finn, K., Coppola, R., Frank, J. A., et al. (1999). Physiological characteristics of capacity constraints in working memory as revealed by functional MRI. *Cerebral Cortex*, 9(1), 20–26.
- Callicott, J. H., Mattay, V. S., Verchinski, B. A., Marenco, S., Egan, M. F., & Weinberger, D. R. (2003). Complexity of prefrontal cortical dysfunction in schizophrenia: More than up or down. *American Journal of Psychiatry*, 160(12), 2209–2215.
- Callicott, J. H., Ramsey, N. F., Tallent, K., Bertolino, A., Knable, M. B., Coppola, R., et al. (1998). Functional magnetic resonance imaging brain mapping in psychiatry: Methodological issues illustrated in a study of working memory in schizophrenia. *Neuropsychopharmacology*, 18(3), 186–196.
- Cannon, T. D. (1997). On the nature and mechanisms of obstetric influences in schizophrenia: A review and synthesis of epidemiologic studies. *International Review of Psychiatry*, 9(4), 387–397.
- Cannon, T. D., Hennah, W., van Erp, T. G., Thompson, P. M., Lonnqvist, J., Huttunen, M., et al. (2005). Association of DISC1/TRAX haplotypes with schizophrenia, reduced prefrontal gray matter, and impaired short- and long-term memory. *Archives of General Psychiatry*, 62(11), 1205–1213.
- Cannon, T. D., Huttunen, M. O., Dahlström, M., Larmo, I., Räsänen, P., & Juriloo, A. (2002). Antipsychotic drug treatment in the prodromal phase of schizophrenia. *Archives of General Psychiatry*, 59(7), 1230–1232.
- Cannon, T. D., Huttunen, M. O., Lonnqvist, J., Tuulio-Henriksson, A., Pirkola, T., Glahn, D., Finkelstein, J., Hietanen, M., Kaprio, J., & Koskenvuo, M. (2000). The inheritance of neuropsychological dysfunction in twins discordant for schizophrenia. *American Journal of Human Genetics*, 67(2), 369–382.
- Cannon, T. D., Kaprio, J., Lonnqvist, J., Huttunen, M., & Koskenvuo, M. (1998). The genetic epidemiology of schizophrenia in a Finnish twin cohort. A population-based modeling study. *Archives of General Psychiatry*, 55(1), 67–74.
- Cannon, T. D., & Rosso, I. M. (2002). Levels of analysis in etiological research on schizophrenia. *Developmental Psychopathology*, 14(3), 653–666.
- Cannon, T. D., Thompson, P. M., van Erp, T. G., Toga, A. W., Poutanen, V. P., Huttunen, M., et al. (2002). Cortex mapping reveals regionally specific patterns of genetic and disease-specific gray-matter deficits in twins discordant for schizophrenia. *Proceedings of the National Academy of Sciences USA*, 99(5), 3228–3233.
- Cannon, T. D., Zorrilla, L. E., Shtasel, D., Gur, R. E., Gur, R. C., Marco, E. J., et al. (1994). Neuropsychological functioning in siblings discordant for schizophrenia and healthy volunteers. *Archives of General Psychiatry*, 51(8), 651–661.
- Chaika, E. (1982). Thought disorder or speech disorder in schizophrenia? *Schizophrenia Bulletin*, 8(4), 587–594.
- Chapman, L. J., & Chapman, J. P. (1973). *Disordered thought in schizophrenia*. New York: Appleton-Century-Crofts.
- Cirillo, M. A., & Seidman, L. J. (2003). Verbal declarative memory dysfunction in schizophrenia: From clinical assessment to genetics and brain mechanisms. *Neuropsychology Review*, 13(2), 43–77.
- Cohen, J. D., Barch, D. M., Carter, C., & Servan-Schreiber, D. (1999). Context-processing deficits in schizophrenia: Converging evidence from three theoretically motivated cognitive tasks. *Journal of Abnormal Psychology*, 108(1), 120–133.
- Cohen, J. D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J., & Smith, E. E. (1997). Temporal dynamics of brain activation during a working memory task. *Nature*, 386(6625), 604–608.
- Cohen, J. D., & Servan-Schreiber, D. (1992). Context, cortex, and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychological Review*, 99(1), 45–77.
- Conrad, A. J., Abebe, T., Austin, R., Forsythe, S., & Scheibel, A. B. (1991). Hippocampal pyramidal cell disarray in schizophrenia as a bilateral phenomenon. *Archives of General Psychiatry*, 48(5), 413–417.

- Conrad, A. J., & Scheibel, A. B. (1987). Schizophrenia and the hippocampus: The embryological hypothesis extended. *Schizophrenia Bulletin*, 13(4), 577–587.
- Costa, A., Strijkers, K., Martin, C., & Thierry, G. (2009). The time course of word retrieval revealed by event-related brain potentials during overt speech. *Proceedings of the National Academy of Sciences USA*, 106(50), 21442–21446.
- Cosway, R., Byrne, M., Clafferty, R., Hodges, A., Grant, E., Abukmeil, S. S., et al. (2000). Neuropsychological change in young people at high risk for schizophrenia: Results from the first two neuropsychological assessments of the Edinburgh High Risk Study. *Psychological Medicine*, 30(5), 1111–1121.
- Critchley, M. (1964). The Neurology of Psychotic Speech. *British Journal of Psychiatry*, 110, 353–364.
- Dauphinais, I. D., DeLisi, L. E., Crow, T. J., Alexandropoulos, K., Colter, N., Tuma, I., & Gershon, E. S. (1990). Reduction in temporal lobe size in siblings with schizophrenia: a magnetic resonance imaging study. *Psychiatry Research*, 35(2), 137–147.
- Ditman, T., & Kuperberg, G.R. (2010). Building coherence: A framework for exploring the breakdown of links across clause boundaries in schizophrenia. *Journal of Neurolinguistics*, 23(3), 254–269.
- Docherty, N. M. (1995). Expressed emotion and language disturbances in parents of stable schizophrenia patients. *Schizophrenia Bulletin*, 21(3), 411–418.
- Docherty, N. M., & Gordinier, S. W. (1999). Immediate memory, attention and communication disturbances in schizophrenia patients and their relatives. *Psychological Medicine*, 29(1), 189–197.
- Dodd, B., & Crosbie, S. (2010). Language and cognition: Evidence from disordered language. In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 604–625). Malden, MA: Blackwell Publishing.
- Dronkers, N.F., Redfern, B.A., & Knight, R.T. (1999). The neural architecture of language disorders. In M. S. Gazzaniga, (Ed.), *The new cognitive neurosciences* (2nd ed., pp. 949–958). Cambridge, MA: MIT Press.
- Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews Neuroscience*, 2(11), 820–829.
- Egan, M. F., Goldberg, T. E., Kolachana, B. S., Callicott, J. H., Mazzanti, C. M., Straub, R. E., et al. (2001). Effect of COMT Val108/158 Met genotype on frontal lobe function and risk for schizophrenia. *Proceedings of the National Academy of Sciences USA*, 98(12), 6917–6922.
- Falkai, P., & Bogerts, B. (1986). Cell loss in the hippocampus of schizophrenics. *European Archives of Psychiatry and Neurological Science*, 236(3), 154–161.
- Farone, S. V., Seidman, L. J., Kremen, W. S., Toomey, R., Peplle, J. R., & Tsuang, M. T. (2000). Neuropsychologic functioning among the nonpsychotic relatives of schizophrenic patients: The effect of genetic loading. *Biological Psychiatry*, 48(2), 120–126.
- Garavan, H., Ross, T. J., Li, S. J., & Stein, E. A. (2000). A parametric manipulation of central executive functioning. *Cerebral Cortex*, 10(6), 585–592.
- Garrod, S., & Pickering, M.J. (2004). Why is conversation so easy? *Trends in Cognitive Science*, 8(1), 8–11.
- Glahn, D. C., Cannon, T. D., Gur, R. E., Ragland, J. D., & Gur, R. C. (2000). Working memory constrains abstraction in schizophrenia. *Biological Psychiatry*, 47(1), 34–42.
- Glahn, D. C., Kim, J., Cohen, M. S., Poutanen, V. P., Theriman, S., Bava, S., et al. (2002). Maintenance and manipulation in spatial working memory: Dissociations in the prefrontal cortex. *Neuroimage*, 17(1), 201–213.
- Glahn, D. C., Ragland, J. D., Abramoff, A., Barrett, J., Laird, A. R., Bearden, C. E., & Velligan, D. I. (2005). Beyond hypofrontality: A quantitative meta-analysis of functional neuroimaging studies of working memory in schizophrenia. *Human Brain Mapping*, 25(1), 60–69.
- Goghari, V. M., Sponheim, S. R., & MacDonald, A. W., III. (2010). The functional neuroanatomy of symptom dimensions in schizophrenia: A qualitative and quantitative review of a persistent question. *Neuroscience and Biobehavioral Reviews*, 34(3), 468–486.
- Gold, B. T., & Buckner, R. L. (2002). Common prefrontal regions coactivate with dissociable posterior regions during controlled semantic and phonological tasks. *Neuron*, 35(4), 803–812.
- Goldberg, T. E., Aloia, M. S., Gourovitch, M. L., Missar, D., Pickar, D., & Weinberger, D. R. (1998). Cognitive substrates of thought disorder, I: The semantic system. *American Journal of Psychiatry*, 155(12), 1671–1676.
- Goldman-Rakic, P. S. (1995). Architecture of the prefrontal cortex and the central executive. *Annals of the New York Academy of Science*, 769, 71–83.
- Goldman-Rakic, P. S., & Selemon, L. D. (1997). Functional and anatomical aspects of prefrontal pathology in schizophrenia. *Schizophrenia Bulletin*, 23(3), 437–458.
- Gottesman, II., & Gould, T. D. (2003). The endophenotype concept in psychiatry: Etymology and strategic intentions. *American Journal of Psychiatry*, 160(4), 636–645.
- Grace, A. A., Moore, H., & O'Donnell, P. (1998). The modulation of corticoaccumbens transmission by limbic afferents and dopamine: A model for the pathophysiology of schizophrenia. *Advances in Pharmacology*, 42, 721–724.
- Gray, J. A., Feldon, J., Rawlins, J. N., Hemsley, D. R., Young, A. M. J., Warburton, E. C., et al. (1991). The neuropsychology of schizophrenia. *Behavioral and Brain Sciences*, 14(1), 1–84.
- Grossman, L. S., & Harrow, M. (1996). Interactive behavior in bipolar manic and schizophrenic patients and its link to thought disorder. *Comprehensive Psychiatry*, 37(4), 245–252.
- Harrow, M., & Quinlan, D. M. (1985). *Disordered thinking and schizophrenic psychopathology*. New York: Gardner Press, Inc.
- Harvey, P. D., & Pedley, M. (1989). Auditory and visual distractibility in schizophrenia. Clinical and medication status correlations. *Schizophrenia Research*, 2(3), 295–300.
- Heckers, S., Rauch, S. L., Goff, D., Savage, C. R., Schacter, D. L., Fischman, A. J., & Alpert, N. M. (1998). Impaired recruitment of the hippocampus during conscious recollection in schizophrenia. *Nature Neuroscience*, 1(4), 318–323.
- Honey, R. C., & Watt, A. (1999). Acquired relational equivalence between contexts and features. *Journal of Experimental Psychology: Animal Behavior Processes*, 25(3), 324–333.
- Horn, L., & Ward, G. (2001). Pragmatics. In F. C. Keil (Ed.), *The MIT encyclopedia of the cognitive sciences (MITECS)* (Vol. 1, pp. 607–632). Cambridge, MA: MIT Press.
- Buttenlocher, P. R. (1979). Synaptic density in human frontal cortex - developmental changes and effects of aging. *Brain Research*, 163(2), 195–205.
- Indefrey, P., & Levelt, W. J. M. (1999). Chapter 59. The neural correlates of language production. In M. S. Gazzaniga (Ed.),

- The new cognitive neurosciences* (2nd ed., pp. 845–866). Cambridge, MA: MIT Press.
- Indefrey, P., & Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition*, 92(1–2), 101–144.
- Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, 30(5), 513–541.
- Jaaro-Peled, H., Ayhan, Y., Pletnikov, M. V., & Sawa, A. (2010). Review of pathological hallmarks of schizophrenia: Comparison of genetic models with patients and nongenetic models. *Schizophrenia Bulletin*, 36(2), 301–313.
- Jeste, D. V., & Lohr, J. B. (1989). Hippocampal pathologic findings in schizophrenia: A morphometric study. *Archives of General Psychiatry*, 46(11), 1019–1024.
- Johnston, M. H., & Holzman, P. S. (1979). *Assessing schizophrenic thinking*. San Francisco, CA: Jossey-Bass.
- Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: an individual-differences perspective. *Psychonomic Bulletin and Review*, 9(4), 637–671.
- Karlsgodt, K. H., Glahn, D. C., van Erp, T. G., Therman, S., Huttunen, M., Manninen, M., et al. (2007). The relationship between performance and fMRI signal during working memory in patients with schizophrenia, unaffected co-twins, and control subjects. *Schizophrenia Research*, 89(1–3), 191–197.
- Karlsgodt, K. H., Sanz, J., van Erp, T.G., Bearden, C.E., Nuechterlein, K.H., & Cannon, T.D. (2009). Re-evaluating dorsolateral prefrontal cortex activation during working memory in schizophrenia. *Schizophrenia Research*, 108(1–3), 143–150.
- Karlsgodt, K. H., van Erp, T. G., Poldrack, R. A., Bearden, C. E., Nuechterlein, K. H., & Cannon T. D. (2008). Diffusion tensor imaging of the superior longitudinal fasciculus and working memory in recent-onset schizophrenia. *Biological Psychiatry*, 63(5), 512–518.
- Kerns, J. G. (2007). Verbal communication impairments and cognitive control components in people with schizophrenia. *Journal of Abnormal Psychology*, 116(2), 279–289.
- Kerns, J. G., & Berenbaum, H. (2002). Cognitive impairments associated with formal thought disorder in people with schizophrenia. *Journal of Abnormal Psychology*, 111(2), 211–224.
- Kerns, J. G., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2004). Prefrontal cortex guides context-appropriate responding during language production. *Neuron*, 43(2), 283–291.
- Kircher, T. T., Liddle, P. F., Brammer, M. J., Williams, S. C., Murray, R. M., & McGuire, P. K. (2001). Neural correlates of formal thought disorder in schizophrenia: Preliminary findings from a functional magnetic resonance imaging study. *Archives of General Psychiatry*, 58(8), 769–774.
- Kostova, M., Passerieu, C., Laurent, J. P., & Hardy-Bayle, M. C. (2003). An electrophysiologic study: Can semantic context processes be mobilized in patients with thought-disordered schizophrenia? *Canadian Journal of Psychiatry*, 48(9), 615–623.
- Kovelman, J. A., & Scheibel, A. B. (1984). A neurohistological correlate of schizophrenia. *Biological Psychiatry*, 19(12), 1601–1621.
- Kumari, V., Fannon, D., Ffytche, D. H., Raveendran, V., Antonova, E., Premkumar, P., et al. (2010). Functional MRI of verbal self-monitoring in schizophrenia: Performance and illness-specific effects. *Schizophrenia Bulletin*, 36(4), 740–755.
- Kuperberg, G. R., Kreher, D. A., & Ditman, T. (2010). What can Event-related Potentials tell us about language, and perhaps even thought, in schizophrenia? *International Journal of Psychophysiology*, 75(2), 66–76.
- Kuperberg, G. R., McGuire, P. K., & David, A. S. (1998). Reduced sensitivity to linguistic context in schizophrenic thought disorder: Evidence from on-line monitoring for words in linguistically anomalous sentences. *Journal of Abnormal Psychology*, 107(3), 423–434.
- Landau, S. M., Lal, R., O'Neil, J. P., Baker, S., & Jagust, W. J. (2009). Striatal dopamine and working memory. *Cerebral Cortex*, 19(2), 445–454.
- Lawrie, S. M., Whalley, H., Kestelman, J. N., Abukmeil, S. S., Byrne, M., Hodges, A., et al. (1999). Magnetic resonance imaging of brain in people at high risk of developing schizophrenia. *Lancet*, 353(9146), 30–33.
- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
- Levelt, W. J. M. (1999). Language production: A blueprint of the speaker. In C. Brown & P. Hagoort (Eds.), *Neurocognition of language* (pp. 94–122). Oxford, England: Oxford University Press.
- Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(1), 1–75.
- Levy, D. L., Coleman, M. J., Sung, H., Ji, F., Matthysse, S., Mendell, N. R., & Titone, D. (2010). The genetic basis of thought disorder and language and communication disturbances in schizophrenia. *Journal of Neurolinguistics*, 23(3), 176.
- Lewis, D. A., & González-Burgos, G. (2008). Neuroplasticity of neocortical circuits in schizophrenia. *Neuropsychopharmacology*, 33(1), 141–165.
- Li, W., Zhou, Y., Jentsch, J. D., Brown, R. A., Tian, X., Ehninger, D., et al. (2007). Specific developmental disruption of disrupted-in-schizophrenia-1 function results in schizophrenia-related phenotypes in mice. *Proceedings of the National Academy of Sciences USA*, 104(46), 18280–18285.
- Liddle, P. F. (1987). The symptoms of chronic schizophrenia. A re-examination of the positive-negative dichotomy. *British Journal of Psychiatry*, 151, 145–151.
- Maher, B. A. (1966). *Principles of psychopathology: An experimental approach*. New York: McGraw-Hill.
- Maher, B. A. (1991). Language and schizophrenia. In J. H. Gruzelier (Ed.), *Neuropsychology, psychophysiology, and information processing: Handbook of schizophrenia*, Vol. 5 (Vol. 5, pp. pp. 437–464). New York: Elsevier Science.
- Maher, B. A., Manschreck, T. C., Woods, B. T., Yurgelun-Todd, D. A., & Tsuang, M. T. (1995). Frontal brain volume and context effects in short-term recall in schizophrenia. *Biological Psychiatry*, 37(3), 144–150.
- Malenka, R. C., Angel, R. W., Hampton, B., & Berger, P. A. (1982). Impaired central error-correcting behavior in schizophrenia. *Archives of General Psychiatry*, 39(1), 101–107.
- Manoach, D. S. (2003). Prefrontal cortex dysfunction during working memory performance in schizophrenia: Reconciling discrepant findings. *Schizophrenia Research*, 60(2–3), 285–298.
- McGrath, J. A., Avramopoulos, D., Lasseter, V. K., Wolyniec, P. S., Fallin, M. D., Liang, K. Y., et al. (2009). Familiality

- of novel factorial dimensions of schizophrenia. *Archives of General Psychiatry*, 66(6), 591–600.
- McNab, F., & Klingberg, T. (2008). Prefrontal cortex and basal ganglia control access to working memory. *Nature Neuroscience*, 11(1), 103–107.
- McNeil, T. F., Cantor-Graae, E., & Weinberger, D. R. (2000). Relationship of obstetric complications and differences in size of brain structures in monozygotic twin pairs discordant for schizophrenia. *American Journal of Psychiatry*, 157(2), 203–212.
- Melinder, M. R., & Barch, D. M. (2003). The influence of a working memory load manipulation on language production in schizophrenia. *Schizophrenia Bulletin*, 29(3), 473–485.
- Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24, 167–202.
- Minzenberg, M. J., Ober, B. A., & Vinogradov, S. (2002). Semantic priming in schizophrenia: A review and synthesis. *Journal of the International Neuropsychology Society*, 8(5), 699–720.
- Nuechterlein, K. H., & Dawson, M. E. (1984). Information processing and attentional functioning in the developmental course of schizophrenic disorders. *Schizophrenia Bulletin*, 10(2), 160–203.
- O'Driscoll, G. A., Florencio, P. S., Gagnon, D., Wolff, A. V., Benkelfat, C., Mikula, L., et al. (2001). Amygdala-hippocampal volume and verbal memory in first-degree relatives of schizophrenic patients. *Psychiatry Research*, 107(2), 75–85.
- Okubo, Y., Suhara, T., Suzuki, K., Kobayashi, K., Inoue, O., Terasaki, O., et al. (1997). Decreased prefrontal dopamine D1 receptors in schizophrenia revealed by PET. *Nature*, 385(6617), 634–636.
- Oltmanns, T. F., & Neale, J. M. (1975). Schizophrenic performance when distractors are present: Attentional deficit or differential task difficulty? *Journal of Abnormal Psychology*, 84(3), 205–209.
- Penn, D. L., Ritchie, M., Francis, J., Combs, D., & Martin, J. (2002). Social perception in schizophrenia: the role of context. *Psychiatry Research*, 109(2), 149–159.
- Peralta, V., Cuesta, M. J., & Farre, C. (1997). Factor structure of symptoms in functional psychoses. *Biological Psychiatry*, 42(9), 806–815.
- Pinkham, A. E., Penn, D. L., Perkins, D. O., & Lieberman, J. (2003). Implications for the neural basis of social cognition for the study of schizophrenia. *American Journal of Psychiatry*, 160(5), 815–824.
- Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). Task context and organization in free recall. *Neuropsychologia*, 47(11), 2158–2163.
- Price, C. J. (2010). The anatomy of language: A review of 100 fMRI studies published in 2009. *Annals of the New York Academy of Sciences*, 1191(1), 62–88.
- Racenstein, J. M., Penn, D., Harrow, M., & Schleser, R. (1999). Thought disorder and psychosocial functioning in schizophrenia: The concurrent and predictive relationships. *Journal of Nervous and Mental Disease*, 187(5), 281–289.
- Ragland, J. D., Yoon, J., Minzenberg, M. J., & Carter, C. S. (2007). Neuroimaging of cognitive disability in schizophrenia: Search for a pathophysiological mechanism. *International Review of Psychiatry*, 19(4), 417–427.
- Rosenberger, G., Kubicki, M., Nestor, P. G., Connor, E., Bushell, G. B., Markant, D., et al. (2008). Age-related deficits in fronto-temporal connections in schizophrenia: a diffusion tensor imaging study. *Schizophrenia Research*, 102(1–3), 181–188.
- Ruff, R. M., Light, R. H., Parker, S. B., & Levin, H. S. (1997). The psychological construct of word fluency. *Brain and Language*, 57(3), 394–405.
- Rutten, B. P., & Mill, J. (2009). Epigenetic mediation of environmental influences in major psychotic disorders. *Schizophrenia Bulletin*, 35(6), 1045–1056.
- Rypma, B., Berger, J. S., & D'Esposito, M. (2002). The influence of working-memory demand and subject performance on prefrontal cortical activity. *Journal of Cognitive Neuroscience*, 14(5), 721–731.
- Sabb, F. W., van Erp, T. G., Hardt, M. E., Dapretto, M., Caplan, R., Cannon, T. D., & Bearden, C. E. (2010). Language network dysfunction as a predictor of outcome in youth at clinical high risk for psychosis. *Schizophrenia Research*, 116(2–3), 173–183.
- Sass, L. A., Gunderson, J. G., Singer, M. T., & Wynne, L. C. (1984). Parental communication deviance and forms of thinking in male schizophrenic offspring. *Journal of Nervous and Mental Disease*, 172(9), 513–520.
- Sassa, Y., Sugiura, M., Jeong, H., Horie, K., Sato, S., & Kawashima, R. (2007). Cortical mechanism of communicative speech production. *Neuroimage*, 37(3), 985–992.
- Saykin, A. J., Shtasel, D. L., Gur, R. E., Kester, D. B., Mozley, L. H., Stafniak, P., & Gur, R. C. (1994). Neuropsychological deficits in neuroleptic naive patients with first-episode schizophrenia. *Archives of General Psychiatry*, 51(2), 124–131.
- Schobel, S. A., Lewandowski, N. M., Corcoran, C. M., Moore, H., Brown, T., Malaspina, D., & Small, S. A. (2009). Differential targeting of the CA1 subfield of the hippocampal formation by schizophrenia and related psychotic disorders. *Archives of General Psychiatry*, 66(9), 938–946.
- Schulze, T. G. (2010). Genetic research into bipolar disorder: The need for a research framework that integrates sophisticated molecular biology and clinically informed phenotype characterization. *Psychiatric Clinics of North America*, 33(1), 67–82.
- Seidman, L. J., Faraone, S. V., Goldstein, J. M., Goodman, J. M., Kremen, W. S., Toomey, R., et al. (1999). Thalamic and amygdala-hippocampal volume reductions in first-degree relatives of patients with schizophrenia: An MRI-based morphometric analysis. *Biological Psychiatry*, 46(7), 941–954.
- Seidman, L. J., Faraone, S. V., Goldstein, J. M., Kremen, W. S., Horton, N. J., Makris, N., et al. (2002). Left hippocampal volume as a vulnerability indicator for schizophrenia: A magnetic resonance imaging morphometric study of nonpsychotic first-degree relatives. *Archives of General Psychiatry*, 59(9), 839–849.
- Seidman, L. J., Giuliano, A. J., Meyer, E. C., Addington, J., Cadenehead, K. S., Cannon, T. D., et al. for the North American Prodrome Longitudinal Study (NAPLS) Group. (2010). Neuropsychology of the prodrome to psychosis in the NAPLS consortium: Relationship to family history and conversion to psychosis. *Archives of General Psychiatry*, 67(6), 578–588.
- Seidman, L. J., Yurgelun-Todd, D., Kremen, W. S., Woods, B. T., Goldstein, J. M., Faraone, S. V., & Tsuang, M. T. (1994). Relationship of prefrontal and temporal lobe MRI measures to neuropsychological performance in chronic schizophrenia. *Biological Psychiatry*, 35(4), 235–246.

- Shenton, M. E., Dickey, C. C., Frumin, M., & McCarley, R. W. (2001). A review of MRI findings in schizophrenia. *Schizophrenia Research*, 49(1–2), 1–52.
- Silver, H., Feldman, P., Bilker, W., & Gur, R.C. (2003). Working memory deficit as a core neuropsychological dysfunction in schizophrenia. *American Journal of Psychiatry*, 160(10), 1809–1816.
- Silverstein, S. M., Kovacs, I., Corry, R., & Valone, C. (2000). Perceptual organization, the disorganization syndrome, and context processing in chronic schizophrenia. *Schizophrenia Research*, 43(1), 11–20.
- Simpson, E. H., Kellendonk, C., & Kandel, E. (2010). A possible role for the striatum in the pathogenesis of the cognitive symptoms of schizophrenia. *Neuron*, 65(5), 585–596.
- Sitnikova, T., Salisbury, D. F., Kuperberg, G., & Holcomb, P. I. (2002). Electrophysiological insights into language processing in schizophrenia. *Psychophysiology*, 39(6), 851–860.
- Spitzer, M., Braun, U., Hermle, L., & Maier, S. (1993). Associative semantic network dysfunction in thought-disordered schizophrenic patients: Direct evidence from indirect semantic priming. *Biological Psychiatry*, 34(12), 864–877.
- Spitzer, M., Weisker, I., Winter, M., Maier, S., Hermle, L., & Maher, B. A. (1994). Semantic and phonological priming in schizophrenia. *Journal of Abnormal Psychology*, 103(3), 485–494.
- Squire, L. R., & Zola, S. M. (1996). Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences USA*, 93(24), 13515–13522.
- Stefanis, N., Frangou, S., Yakeley, J., Sharma, T., O'Connell, P., Morgan, K., et al. (1999). Hippocampal volume reduction in schizophrenia: effects of genetic risk and pregnancy and birth complications. *Biological Psychiatry*, 46(5), 697–702.
- Straub, R. E., Lipska, B. K., Egan, M. F., Goldberg, T. E., Callicott, J. H., Mayhew, M. B., et al. (2007). Allelic variation in GAD1 (GAD67) is associated with schizophrenia and influences cortical function and gene expression. *Molecular Psychiatry*, 12(9), 854–869.
- Sun, D., Stuart, G.W., Jenkins, M., Wood, S.J., McGorry, P. D., Velakoulis, D., et al. (2009). Brain surface contraction mapped in first-episode schizophrenia: A longitudinal magnetic resonance imaging study. *Molecular Psychiatry*, 14(10), 976–986.
- Sun, D., van Erp, T.G., Thompson, P.M., Bearden, C.E., Daley, M., Kushan, L., et al. (2009). Elucidating a magnetic resonance imaging-based neuroanatomic biomarker for psychosis: Classification analysis using probabilistic brain atlas and machine learning algorithms. *Biological Psychiatry*, 66(11), 1055–1060.
- Tan, H. Y., Callicott, J. H., & Weinberger, D. R. (2009). Prefrontal cognitive systems in schizophrenia: Towards human genetic brain mechanisms. *Cognitive Neuropsychiatry*, 14(4–5), 277–298.
- Tandon, R., Keshavan, M. S., & Nasrallah, H. A. (2008). Schizophrenia, “just the facts” what we know in 2008. 2. Epidemiology and etiology. *Schizophrenia Research*, 102(1–3), 1–18.
- Titone, D., Levy, D. L., & Holzman, P. S. (2000). Contextual insensitivity in schizophrenic language processing: Evidence from lexical ambiguity. *Journal of Abnormal Psychology*, 109(4), 761–767.
- Tsuang, M. T., & Faraone, S. V. (1999). The concept of target features in schizophrenia research. *Acta Psychiatrica Scandinavica*, 99(395 Suppl), 2–11.
- Tsuang, M. T., Stone, W. S., & Faraone, S. V. (1999). Schizophrenia: A review of genetic studies. *Harvard Review of Psychiatry*, 7(4), 185–207.
- van Erp, T. G., Lesh, T. A., Knowlton, B. J., Bearden, C. E., Hardt, M., Karlsgodt, K. H., et al. (2008). Remember and know judgments during recognition in chronic schizophrenia. *Schizophrenia Research*, 100(1–3), 181–190.
- Van Erp, T. G., Saleh, P. A., Rosso, I. M., Huttunen, M., Lonnqvist, J., Pirkola, T., et al. (2002). Contributions of genetic risk and fetal hypoxia to hippocampal volume in patients with schizophrenia or schizoaffective disorder, their unaffected siblings, and healthy unrelated volunteers. *American Journal of Psychiatry*, 159(9), 1514–1520.
- Van Snellenberg, J. X., Torres, I. J., & Thornton, A. E. (2006). Functional neuroimaging of working memory in schizophrenia: Task performance as a moderating variable. *Neuropsychology*, 20(5), 497–510.
- van Veen, V., & Carter, C. S. (2006). Error detection, correction, and prevention in the brain: a brief review of data and theories. *Clinical EEG and Neurosciences*, 37(4), 330–335.
- Vargha-Khadem, F., Gadian, D. G., Watkins, K. E., Connelly, A., Van Paesschen, W., & Mishkin, M. (1997). Differential effects of early hippocampal pathology on episodic and semantic memory. *Science*, 277(5324), 376–380.
- Velakoulis, D., Pantelis, C., McGorry, P. D., Dudgeon, P., Brewer, W., Cook, M., et al. (1999). Hippocampal volume in first-episode psychoses and chronic schizophrenia: A high-resolution magnetic resonance imaging study. *Archives of General Psychiatry*, 56(2), 133–141.
- Velakoulis, D., Wood, S. J., Wong, M. T., McGorry, P. D., Yung, A., Phillips, L., et al. (2006). Hippocampal and amygdala volumes according to psychosis stage and diagnosis: A magnetic resonance imaging study of chronic schizophrenia, first-episode psychosis, and ultra-high-risk individuals. *Archives of General Psychiatry*, 63(2), 139–149.
- Wahlberg, K. E., Wynne, L. C., Oja, H., Keskitalo, P., Anais-Tanner, H., Koistinen, P., et al. (2000). Thought disorder index of Finnish adoptees and communication deviance of their adoptive parents. *Psychological Medicine*, 30(1), 127–136.
- Weisbrod, M., Maier, S., Harig, S., Himmelsbach, U., & Spitzer, M. (1998). Lateralised semantic and indirect semantic priming effects in people with schizophrenia. *British Journal of Psychiatry*, 172, 142–146.
- Wirshing, D. A., Marshall, B. D., Green, M. F., Mintz, J., Marder, S. R., & Wirshing, W. C. (1999). Risperidone in treatment-refractory schizophrenia. *American Journal of Psychiatry*, 156(9), 1374–1379.
- Wood, J. N. (2003). Social cognition and the prefrontal cortex. *Behavioral and Cognitive Neuroscience Reviews*, 2(2), 97–114.
- Wood, S. J., Kennedy, D., Phillips, L. J., Seal, M. L., Yücel, M., Nelson, B., et al. (2010). Hippocampal pathology in individuals at ultra-high risk for psychosis: A multimodal magnetic resonance study. *Neuroimage*, 52(1), 62–68.
- Yurgelun-Todd, D. A., Waternaux, C. M., Cohen, B. M., Gruber, S. A., English, C. D., & Renshaw, P. F. (1996). Functional

- magnetic resonance imaging of schizophrenic patients and comparison subjects during word production. *American Journal of Psychiatry*, 153(2), 200–205.
- Zaidel, D. W., Esiri, M. M., & Harrison, P. J. (1997). Size, shape, and orientation of neurons in the left and right hippocampus: Investigation of normal asymmetries and alterations in schizophrenia. *American Journal of Psychiatry*, 154(6), 812–818.
- Zola, S. M., & Squire, L. R. (2001). Relationship between magnitude of damage to the hippocampus and impaired recognition memory in monkeys. *Hippocampus*, 11(2), 92–98.

This page intentionally left blank

PART 7

Thinking in Practice

This page intentionally left blank

Scientific Thinking and Reasoning

Kevin N. Dunbar and David Klahr

Abstract

Scientific thinking refers to both thinking about the *content* of science and the set of *reasoning processes* that permeate the field of science: induction, deduction, experimental design, causal reasoning, concept formation, hypothesis testing, and so on. Here we cover both the history of research on scientific thinking and the different approaches that have been used, highlighting common themes that have emerged over the past 50 years of research. Future research will focus on the collaborative aspects of scientific thinking, on effective methods for teaching science, and on the neural underpinnings of the scientific mind.

Keywords: scientific reasoning, causal reasoning, hypothesis testing, analogical reasoning in science, problem solving, conceptual change, computational modeling, constructivism, science education, cognitive development, educational neuroscience

There is no unitary activity called “scientific discovery”; there are activities of designing experiments, gathering data, inventing and developing observational instruments, formulating and modifying theories, deducing consequences from theories, making predictions from theories, testing theories, inducing regularities and invariants from data, discovering theoretical constructs, and others.

—Simon, Langley, & Bradshaw, 1981, p. 2

What Is Scientific Thinking and Reasoning?

There are two kinds of thinking we call “scientific.” The first, and most obvious, is thinking about the *content* of science. People are engaged in scientific thinking when they are reasoning about such entities and processes as force, mass, energy, equilibrium, magnetism, atoms, photosynthesis, radiation, geology, or astrophysics (and, of course, cognitive psychology!). The second kind of scientific thinking includes the set of *reasoning processes* that permeate

the field of science: induction, deduction, experimental design, causal reasoning, concept formation, hypothesis testing, and so on. However, these reasoning processes are not unique to scientific thinking: They are the very same processes involved in everyday thinking. As Einstein put it:

The scientific way of forming concepts differs from that which we use in our daily life, not basically, but merely in the more precise definition of concepts and conclusions; more painstaking and systematic choice of experimental material, and greater logical economy. (The Common Language of Science, 1941, reprinted in Einstein, 1950, p. 98)

Nearly 40 years after Einstein’s remarkably insightful statement, Francis Crick offered a similar perspective: that great discoveries in science result not from extraordinary mental processes, but rather from rather common ones. The greatness of the discovery lies in the thing discovered.

I think what needs to be emphasized about the discovery of the double helix is that the path to it was, scientifically speaking, fairly commonplace. What was important was *not the way it was discovered*, but the object discovered—the structure of DNA itself. (Crick , 1988, p. 67; emphasis added)

Under this view, scientific thinking involves the same general-purpose cognitive processes—such as induction, deduction, analogy, problem solving, and causal reasoning—that humans apply in non-scientific domains. These processes are covered in several different chapters of this handbook: Rips, Smith, & Medin, Chapter 11 on induction; Evans, Chapter 8 on deduction; Holyoak, Chapter 13 on analogy; Bassok & Novick, Chapter 21 on problem solving; and Cheng & Buehner, Chapter 12 on causality. One might question the claim that the highly specialized procedures associated with doing science in the “real world” can be understood by investigating the thinking processes used in laboratory studies of the sort described in this volume. However, when the focus is on major scientific breakthroughs, rather than on the more routine, incremental progress in a field, the psychology of problem solving provides a rich source of ideas about how such discoveries might occur. As Simon and his colleagues put it:

It is understandable, if ironic, that ‘normal’ science fits . . . the description of expert problem solving, while ‘revolutionary’ science fits the description of problem solving by novices. It is understandable because scientific activity, particularly at the revolutionary end of the continuum, is concerned with the discovery of new truths, not with the application of truths that are already well-known . . . it is basically a journey into unmapped terrain. Consequently, it is mainly characterized, as is novice problem solving, by trial-and-error search. The search may be highly selective—but it reaches its goal only after many halts, turnings, and back-trackings. (Simon, Langley, & Bradshaw, 1981, p. 5)

The research literature on scientific thinking can be roughly categorized according to the two types of scientific thinking listed in the opening paragraph of this chapter: (1) One category focuses on thinking that directly involves scientific *content*. Such research ranges from studies of young children reasoning about the sun-moon-earth system (Vosniadou & Brewer, 1992) to college students reasoning about chemical equilibrium (Davenport, Yaron, Klahr, &

Koedinger, 2008), to research that investigates collaborative problem solving by world-class researchers in real-world molecular biology labs (Dunbar, 1995). (2) The other category focuses on “general” cognitive processes, but it tends to do so by analyzing people’s problem-solving behavior when they are presented with relatively complex situations that involve the integration and coordination of several different types of processes, and that are designed to capture some essential features of “real-world” science in the psychology laboratory (Bruner, Goodnow, & Austin, 1956; Klahr & Dunbar, 1988; Mynatt, Doherty, & Tweney, 1977).

There are a number of overlapping research traditions that have been used to investigate scientific thinking. We will cover both the history of research on scientific thinking and the different approaches that have been used, highlighting common themes that have emerged over the past 50 years of research.

A Brief History of Research on Scientific Thinking

Science is often considered one of the hallmarks of the human species, along with art and literature. Illuminating the thought processes used in science thus reveal key aspects of the human mind. The thought processes underlying scientific thinking have fascinated both scientists and nonscientists because the products of science have transformed our world and because the process of discovery is shrouded in mystery. Scientists talk of the chance discovery, the flash of insight, the years of perspiration, and the voyage of discovery. These images of science have helped make the mental processes underlying the discovery process intriguing to cognitive scientists as they attempt to uncover what really goes on inside the scientific mind and how scientists really think. Furthermore, the possibilities that scientists can be taught to think better by avoiding mistakes that have been clearly identified in research on scientific thinking, and that their scientific process could be partially automated, makes scientific thinking a topic of enduring interest.

The cognitive processes underlying scientific discovery and day-to-day scientific thinking have been a topic of intense scrutiny and speculation for almost 400 years (e.g., Bacon, 1620; Galilei, 1638; Klahr, 2000; Tweney, Doherty, & Mynatt, 1981). Understanding the nature of scientific thinking has been a central issue not only for our understanding of science but also for our understanding of what it is to be human. Bacon’s *Novum Organum* in 1620

sketched out some of the key features of the ways that experiments are designed and data interpreted. Over the ensuing 400 years philosophers and scientists vigorously debated about the appropriate methods that scientists should use (see Giere, 1993). These debates over the appropriate methods for science typically resulted in the espousal of a particular type of reasoning method, such as induction or deduction. It was not until the Gestalt psychologists began working on the nature of human problem solving, during the 1940s, that experimental psychologists began to investigate the cognitive processes underlying scientific thinking and reasoning.

The Gestalt psychologist Max Wertheimer pioneered the investigation of scientific thinking (of the first type described earlier: thinking about scientific content) in his landmark book *Productive Thinking* (Wertheimer, 1945). Wertheimer spent a considerable amount of time corresponding with Albert Einstein, attempting to discover how Einstein generated the concept of relativity. Wertheimer argued that Einstein had to overcome the structure of Newtonian physics at each step in his theorizing, and the ways that Einstein actually achieved this restructuring were articulated in terms of Gestalt theories. (For a recent and different account of how Einstein made his discovery, see Galison, 2003.) We will see later how this process of overcoming alternative theories is an obstacle that both scientists and nonscientists need to deal with when evaluating and theorizing about the world.

One of the first investigations of scientific thinking of the second type (i.e., collections of general-purpose processes operating on complex, abstract, components of scientific thought) was carried out by Jerome Bruner and his colleagues at Harvard (Bruner et al., 1956). They argued that a key activity engaged in by scientists is to determine whether a particular instance is a member of a category. For example, a scientist might want to discover which substances undergo fission when bombarded by neutrons and which substances do not. Here, scientists have to discover the attributes that make a substance undergo fission. Bruner et al. saw scientific thinking as the testing of hypotheses and the collecting of data with the end goal of determining whether something is a member of a category. They invented a paradigm where people were required to formulate hypotheses and collect data that test their hypotheses. In one type of experiment, the participants were shown a card such as one with two borders and three green triangles. The participants

were asked to determine the concept that this card represented by choosing other cards and getting feedback from the experimenter as to whether the chosen card was an example of the concept. In this case the participant may have thought that the concept was green and chosen a card with two green squares and one border. If the underlying concept was green, then the experimenter would say that the card was an example of the concept. In terms of scientific thinking, choosing a new card is akin to conducting an experiment, and the feedback from the experimenter is similar to knowing whether a hypothesis is confirmed or disconfirmed. Using this approach, Bruner et al. identified a number of strategies that people use to formulate and test hypotheses. They found that a key factor determining which hypothesis-testing strategy that people use is the amount of memory capacity that the strategy takes up (see also Morrison & Knowlton, Chapter 6; Medin et al., Chapter 11). Another key factor that they discovered was that it was much more difficult for people to discover negative concepts (e.g., not blue) than positive concepts (e.g., blue). Although Bruner et al.'s research is most commonly viewed as work on concepts, they saw their work as uncovering a key component of scientific thinking.

A second early line of research on scientific thinking was developed by Peter Wason and his colleagues (Wason, 1968). Like Bruner et al., Wason saw a key component of scientific thinking as being the testing of hypotheses. Whereas Bruner et al. focused on the different types of strategies that people use to formulate hypotheses, Wason focused on whether people adopt a strategy of trying to confirm or disconfirm their hypotheses. Using Popper's (1959) theory that scientists should try and falsify rather than confirm their hypotheses, Wason devised a deceptively simple task in which participants were given three numbers, such as 2-4-6, and were asked to discover the rule underlying the three numbers. Participants were asked to generate other triads of numbers and the experimenter would tell the participant whether the triad was consistent or inconsistent with the rule. They were told that when they were sure they knew what the rule was they should state it. Most participants began the experiment by thinking that the rule was even numbers increasing by 2. They then attempted to confirm their hypothesis by generating a triad like 8-10-12, then 14-16-18. These triads are consistent with the rule and the participants were told yes, that the triads were indeed consistent with the rule. However, when

they proposed the rule—even numbers increasing by 2—they were told that the rule was incorrect. The correct rule was numbers of increasing magnitude! From this research, Wason concluded that people try to confirm their hypotheses, whereas normatively speaking, they should try to disconfirm their hypotheses. One implication of this research is that confirmation bias is not just restricted to scientists but is a general human tendency.

It was not until the 1970s that a general account of scientific reasoning was proposed. Herbert Simon, often in collaboration with Allan Newell, proposed that scientific thinking is a form of problem solving. He proposed that problem solving is a search in a problem space. Newell and Simon's theory of problem solving is discussed in many places in this handbook, usually in the context of specific problems (see especially Bassok & Novick, Chapter 21). Herbert Simon, however, devoted considerable time to understanding many different scientific discoveries and scientific reasoning processes. The common thread in his research was that scientific thinking and discovery is not a mysterious magical process but a process of problem solving in which clear heuristics are used. Simon's goal was to articulate the heuristics that scientists use in their research at a fine-grained level. By constructing computer programs that simulated the process of several major scientific discoveries, Simon and colleagues were able to articulate the specific computations that scientists could have used in making those discoveries (Langley, Simon, Bradshaw, & Zytkow, 1987; see section on "Computational Approaches to Scientific Thinking"). Particularly influential was Simon and Lea's (1974) work demonstrating that concept formation and induction consist of a search in two problem spaces: a space of instances and a space of rules. This idea has influenced problem-solving accounts of scientific thinking that will be discussed in the next section.

Overall, the work of Bruner, Wason, and Simon laid the foundations for contemporary research on scientific thinking. Early research on scientific thinking is summarized in Tweney, Doherty and Mynatt's 1981 book *On Scientific Thinking*, where they sketched out many of the themes that have dominated research on scientific thinking over the past few decades. Other more recent books such as *Cognitive Models of Science* (Giere, 1993), *Exploring Science* (Klahr, 2000), *Cognitive Basis of Science* (Carruthers, Stich, & Siegal, 2002), and *New Directions in Scientific and Technical Thinking* (Gorman, Kincannon, Gooding,

& Tweney, 2004) provide detailed analyses of different aspects of scientific discovery. Another important collection is Vosniadou's handbook on conceptual change research (Vosniadou, 2008). In this chapter, we discuss the main approaches that have been used to investigate scientific thinking.

How does one go about investigating the many different aspects of scientific thinking? One common approach to the study of the scientific mind has been to investigate several key aspects of scientific thinking using abstract tasks designed to mimic some essential characteristics of "real-world" science. There have been numerous methodologies that have been used to analyze the genesis of scientific concepts, theories, hypotheses, and experiments. Researchers have used experiments, verbal protocols, computer programs, and analyzed particular scientific discoveries. A more recent development has been to increase the ecological validity of such research by investigating scientists as they reason "live" (in vivo studies of scientific thinking) in their own laboratories (Dunbar, 1995, 2002). From a "Thinking and Reasoning" standpoint the major aspects of scientific thinking that have been most actively investigated are problem solving, analogical reasoning, hypothesis testing, conceptual change, collaborative reasoning, inductive reasoning, and deductive reasoning.

SCIENTIFIC THINKING AS PROBLEM SOLVING

One of the primary goals of accounts of scientific thinking has been to provide an overarching framework to understand the scientific mind. One framework that has had a great influence in cognitive science is that scientific thinking and scientific discovery can be conceived as a form of problem solving. As noted in the opening section of this chapter, Simon (1977; Simon, Langley, & Bradshaw, 1981) argued that both scientific thinking in general and problem solving in particular could be thought of as a search in a problem space. A problem space consists of all the possible states of a problem and all the operations that a problem solver can use to get from one state to the next. According to this view, by characterizing the types of representations and procedures that people use to get from one state to another it is possible to understand scientific thinking. Thus, scientific thinking can be characterized as a search in various problem spaces (Simon, 1977). Simon investigated a number of scientific discoveries by bringing participants into the laboratory, providing the participants with the data that a scientist had access to, and getting the participants to reason

about the data and rediscover a scientific concept. He then analyzed the verbal protocols that participants generated and mapped out the types of problem spaces that the participants search in (e.g., Qin & Simon, 1990). Kulkarni and Simon (1988) used a more historical approach to uncover the problem-solving heuristics that Krebs used in his discovery of the urea cycle. Kulkarni and Simon analyzed Krebs's diaries and proposed a set of problem-solving heuristics that he used in his research. They then built a computer program incorporating the heuristics and biological knowledge that Krebs had before he made his discoveries. Of particular importance are the search heuristics that the program uses, which include experimental proposal heuristics and data interpretation heuristics. A key heuristic was an unusualness heuristic that focused on unusual findings, which guided search through a space of theories and a space of experiments.

Klahr and Dunbar (1988) extended the search in a problem space approach and proposed that scientific thinking can be thought of as a search through two related spaces: an hypothesis space and an experiment space. Each problem space that a scientist uses will have its own types of representations and operators used to change the representations. Search in the hypothesis space constrains search in the experiment space. Klahr and Dunbar found that some participants move from the hypothesis space to the experiment space, whereas others move from the experiment space to the hypothesis space. These different types of searches lead to the proposal of different types of hypotheses and experiments. More recent work has extended the dual-space approach to include alternative problem-solving spaces, including those for data, instrumentation, and domain-specific knowledge (Klahr & Simon, 1999; Schunn & Klahr, 1995, 1996).

SCIENTIFIC THINKING AS HYPOTHESIS TESTING

Many researchers have regarded testing specific hypotheses predicted by theories as one of the key attributes of scientific thinking. Hypothesis testing is the process of evaluating a proposition by collecting evidence regarding its truth. Experimental cognitive research on scientific thinking that specifically examines this issue has tended to fall into two broad classes of investigations. The first class is concerned with the types of reasoning that lead scientists astray, thus blocking scientific ingenuity. A large amount of research has been conducted on the potentially

faulty reasoning strategies that both participants in experiments and scientists use, such as considering only one favored hypothesis at a time and how this prevents the scientists from making discoveries. The second class is concerned with uncovering the mental processes underlying the generation of new scientific hypotheses and concepts. This research has tended to focus on the use of analogy and imagery in science, as well as the use of specific types of problem-solving heuristics.

Turning first to investigations of what diminishes scientific creativity, philosophers, historians, and experimental psychologists have devoted a considerable amount of research to "confirmation bias." This occurs when scientists only consider one hypothesis (typically the favored hypothesis) and ignore other alternative hypotheses or potentially relevant hypotheses. This important phenomenon can distort the design of experiments, formulation of theories, and interpretation of data. Beginning with the work of Wason (1968) and as discussed earlier, researchers have repeatedly shown that when participants are asked to design an experiment to test a hypothesis they will predominantly design experiments that they think will yield results consistent with the hypothesis. Using the 2-4-6 task mentioned earlier, Klayman and Ha (1987) showed that in situations where one's hypothesis is likely to be confirmed, seeking confirmation is a normatively incorrect strategy, whereas when the probability of confirming one's hypothesis is low, then attempting to confirm one's hypothesis can be an appropriate strategy. Historical analyses by Tweney (1989), concerning the way that Faraday made his discoveries, and experiments investigating people testing hypotheses, have revealed that people use a confirm early, disconfirm late strategy: When people initially generate or are given hypotheses, they try and gather evidence that is consistent with the hypothesis. Once enough evidence has been gathered, then people attempt to find the boundaries of their hypothesis and often try to disconfirm their hypotheses.

In an interesting variant on the confirmation bias paradigm, Gorman (1989) showed that when participants are told that there is the possibility of error in the data that they receive, participants assume that any data that are inconsistent with their favored hypothesis are due to error. Thus, the possibility of error "insulates" hypotheses against disconfirmation. This intriguing hypothesis has not been confirmed by other researchers (Penner & Klahr, 1996),

but it is an intriguing hypothesis that warrants further investigation.

Confirmation bias is very difficult to overcome. Even when participants are asked to consider alternate hypotheses, they will often fail to conduct experiments that could potentially disconfirm their hypothesis. Tweney and his colleagues provide an excellent overview of this phenomenon in their classic monograph *On Scientific Thinking* (1981). The precise reasons for this type of block are still widely debated. Researchers such as Michael Doherty have argued that working memory limitations make it difficult for people to consider more than one hypothesis. Consistent with this view, Dunbar and Sussman (1995) have shown that when participants are asked to hold irrelevant items in working memory while testing hypotheses, the participants will be unable to switch hypotheses in the face of inconsistent evidence. While working memory limitations are involved in the phenomenon of confirmation bias, even groups of scientists can also display confirmation bias. For example, the controversy over cold fusion is an example of confirmation bias. Here, large groups of scientists had other hypotheses available to explain their data yet maintained their hypotheses in the face of other more standard alternative hypotheses. Mitroff (1974) provides some interesting examples of NASA scientists demonstrating confirmation bias, which highlight the roles of commitment and motivation in this process. See also MacPherson and Stanovich (2007) for specific strategies that can be used to overcome confirmation bias.

Causal Thinking in Science

Much of scientific thinking and scientific theory building pertains to the development of causal models between variables of interest. For example, do vaccines cause illnesses? Do carbon dioxide emissions cause global warming? Does water on a planet indicate that there is life on the planet? Scientists and nonscientists alike are constantly bombarded with statements regarding the causal relationship between such variables. How does one evaluate the status of such claims? What kinds of data are informative? How do scientists and nonscientists deal with data that are inconsistent with their theory?

A central issue in the causal reasoning literature, one that is directly relevant to scientific thinking, is the extent to which scientists and nonscientists alike are governed by the search for causal mechanisms (i.e., how a variable works) versus the search for statistical data (i.e., how often variables co-occur).

This dichotomy can be boiled down to the search for qualitative versus quantitative information about the paradigm the scientist is investigating. Researchers from a number of cognitive psychology laboratories have found that people prefer to gather more information about an underlying mechanism than covariation between a cause and an effect (e.g., Ahn, Kalish, Medin, & Gelman, 1995). That is, the predominant strategy that students in simulations of scientific thinking use is to gather as much information as possible about how the objects under investigation work, rather than collecting large amounts of quantitative data to determine whether the observations hold across multiple samples. These findings suggest that a central component of scientific thinking may be to formulate explicit mechanistic causal models of scientific events.

One type of situation in which causal reasoning has been observed extensively is when scientists obtain unexpected findings. Both historical and naturalistic research has revealed that reasoning causally about unexpected findings plays a central role in science. Indeed, scientists themselves frequently state that a finding was due to chance or was unexpected. Given that claims of unexpected findings are such a frequent component of scientists' autobiographies and interviews in the media, Dunbar (1995, 1997, 1999; Dunbar & Fugelsang, 2005; Fugelsang, Stein, Green, & Dunbar, 2004) decided to investigate the ways that scientists deal with unexpected findings. In 1991–1992 Dunbar spent 1 year in three molecular biology laboratories and one immunology laboratory at a prestigious U.S. university. He used the weekly laboratory meeting as a source of data on scientific discovery and scientific reasoning. (He termed this type of study "in vivo" cognition.) When he looked at the types of findings that the scientists made, he found that over 50% of the findings were unexpected and that these scientists had evolved a number of effective strategies for dealing with such findings. One clear strategy was to reason causally about the findings: Scientists attempted to build causal models of their unexpected findings. This causal model building results in the extensive use of collaborative reasoning, analogical reasoning, and problem-solving heuristics (Dunbar, 1997, 2001).

Many of the key unexpected findings that scientists reasoned about in the in vivo studies of scientific thinking were inconsistent with the scientists' preexisting causal models. A laboratory equivalent of the biology labs involved creating a situation in

which students obtained unexpected findings that were inconsistent with their preexisting theories. Dunbar and Fugelsang (2005) examined this issue by creating a scientific causal thinking simulation where experimental outcomes were either expected or unexpected. Dunbar (1995) has called the study of people reasoning in a cognitive laboratory “*in vitro*” cognition. These investigators found that students spent considerably more time reasoning about unexpected findings than expected findings. In addition, when assessing the overall degree to which their hypothesis was supported or refuted, participants spent the majority of their time considering unexpected findings. An analysis of participants’ verbal protocols indicates that much of this extra time was spent formulating causal models for the unexpected findings. Similarly, scientists spend more time considering unexpected than expected findings, and this time is devoted to building causal models (Dunbar & Fugelsang, 2004).

Scientists know that unexpected findings occur often, and they have developed many strategies to take advantage of their unexpected findings. One of the most important places that they anticipate the unexpected is in designing experiments (Baker & Dunbar, 2000). They build different causal models of their experiments incorporating many conditions and controls. These multiple conditions and controls allow unknown mechanisms to manifest themselves. Thus, rather than being the victims of the unexpected, they create opportunities for unexpected events to occur, and once these events do occur, they have causal models that allow them to determine exactly where in the causal chain their unexpected finding arose. The results of these *in vivo* and *in vitro* studies all point to a more complex and nuanced account of how scientists and nonscientists alike test and evaluate hypotheses about theories.

The Roles of Inductive, Abductive, and Deductive Thinking in Science

One of the most basic characteristics of science is that scientists assume that the universe that we live in follows predictable rules. Scientists reason using a variety of different strategies to make new scientific discoveries. Three frequently used types of reasoning strategies that scientists use are inductive, abductive, and deductive reasoning. In the case of inductive reasoning, a scientist may observe a series of events and try to discover a rule that governs the event. Once a rule is discovered, scientists can extrapolate from the rule to formulate theories of observed and yet-to-be-observed phenomena. One example is the discovery

using inductive reasoning that a certain type of bacterium is a cause of many ulcers (Thagard, 1999). In a fascinating series of articles, Thagard documented the reasoning processes that Marshall and Warren went through in proposing this novel hypothesis. One key reasoning process was the use of induction by generalization. Marshall and Warren noted that almost all patients with gastric entritis had a spiral bacterium in their stomachs, and he formed the generalization that this bacterium is the cause of stomach ulcers. There are numerous other examples of induction by generalization in science, such as Tycho De Brea’s induction about the motion of planets from his observations, Dalton’s use of induction in chemistry, and the discovery of prions as the source of mad cow disease. Many theories of induction have used scientific discovery and reasoning as examples of this important reasoning process.

Another common type of inductive reasoning is to map a feature of one member of a category to another member of a category. This is called categorical induction. This type of induction is a way of projecting a known property of one item onto another item that is from the same category. Thus, knowing that the Rous Sarcoma virus is a retrovirus that uses RNA rather than DNA, a biologist might assume that another virus that is thought to be a retrovirus also uses RNA rather than DNA. While research on this type of induction typically has not been discussed in accounts of scientific thinking, this type of induction is common in science. For an influential contribution to this literature, see Smith, Shafir, and Osherson (1993), and for reviews of this literature see Heit (2000) and Medin et al. (Chapter 11).

While less commonly mentioned than inductive reasoning, abductive reasoning is an important form of reasoning that scientists use when they are seeking to propose explanations for events such as unexpected findings (see Lombrozo, Chapter 14; Magnani, et al., 2010). In Figure 35.1, taken from King (2011), the differences between inductive, abductive, and deductive thinking are highlighted. In the case of abduction, the reasoner attempts to generate explanations of the form “if situation X had occurred, could it have produced the current evidence I am attempting to interpret?” (For an interesting analysis of abductive reasoning see the brief paper by Klahr & Masnick, 2001). Of course, as in classical induction, such reasoning may produce a plausible account that is still not the correct one. However, abduction does involve the generation of new knowledge, and is thus also related to research on creativity.



Like humans, robots can use various methods of reasoning. The methods may or may not be sound, but they provide ways to form hypotheses and suggest experiments that can be performed to test those hypotheses.

Fig. 35.1 The different processes underlying inductive, abductive, and deductive reasoning in science. (Figure reproduced from King 2011.)

Turning now to deductive thinking, many thinking processes that scientists adhere to follow traditional rules of deductive logic. These processes correspond to those conditions in which a hypothesis may lead to, or is deducible to, a conclusion. Though they are not always phrased in syllogistic form, deductive arguments can be phrased as “syllogisms,” or as brief, mathematical statements in which the premises lead to the conclusion. Deductive reasoning is an extremely important aspect of scientific thinking because it underlies a large component of how scientists conduct their research. By looking at many scientific discoveries, we can often see that deductive reasoning is at work. Deductive reasoning statements all contain information or rules that state an assumption about how the world works, as well as a conclusion that would necessarily follow from the rule. Numerous discoveries in physics such as the discovery of dark matter by Vera Rubin are

based on deductions. In the dark matter case, Rubin measured galactic rotation curves and based on the differences between the predicted and observed angular motions of galaxies she deduced that the structure of the universe was uneven. This led her to propose that dark matter existed. In contemporary physics the CERN Large Hadron Collider is being used to search for the Higgs Boson. The Higgs Boson is a deductive prediction from contemporary physics. If the Higgs Boson is not found, it may lead to a radical revision of the nature of physics and a new understanding of mass (Hecht, 2011).

The Roles of Analogy in Scientific Thinking

One of the most widely mentioned reasoning processes used in science is analogy. Scientists use analogies to form a bridge between what they already know and what they are trying to explain, understand, or discover. In fact, many scientists have claimed that

the making of certain analogies was instrumental in their making a scientific discovery, and almost all scientific autobiographies and biographies feature one particular analogy that is discussed in depth. Coupled with the fact that there has been an enormous research program on analogical thinking and reasoning (see Holyoak, Chapter 13), we now have a number of models and theories of analogical reasoning that suggest how analogy can play a role in scientific discovery (see Gentner, Holyoak, & Kokinov, 2001). By analyzing several major discoveries in the history of science, Thagard and Croft (1999), Nersessian (1999, 2008), and Gentner and Jeziorski (1993) have all shown that analogical reasoning is a key aspect of scientific discovery.

Traditional accounts of analogy distinguish between two components of analogical reasoning: the target and the source (Holyoak, Chapter 13; Gentner, 2010). The target is the concept or problem that a scientist is attempting to explain or solve. The source is another piece of knowledge that the scientist uses to understand the target or to explain the target to others. What the scientist does when he or she makes an analogy is to map features of the source onto features of the target. By mapping the features of the source onto the target, new features of the target may be discovered, or the features of the target may be rearranged so that a new concept is invented and a scientific discovery is made. For example, a common analogy that is used with computers is to describe a harmful piece of software as a computer virus. Once a piece of software is called a virus, people can map features of biological viruses, such as that it is small, spreads easily, self-replicates using a host, and causes damage. People not only map individual features of the source onto the target but also the systems of relations. For example, if a computer virus is similar to a biological virus, then an immune system can be created on computers that can protect computers from future variants of a virus. One of the reasons that scientific analogy is so powerful is that it can generate new knowledge, such as the creation of a computational immune system having many of the features of a real biological immune system. This analogy also leads to predictions that there will be newer computer viruses that are the computational equivalent of retroviruses, lacking DNA, or standard instructions, that will elude the computational immune system.

The process of making an analogy involves a number of key steps: retrieval of a source from memory, aligning the features of the source with those of the

target, mapping features of the source onto those of the target, and possibly making new inferences about the target. Scientific discoveries are made when the source highlights a hitherto unknown feature of the target or restructures the target into a new set of relations. Interestingly, research on analogy has shown that participants do not easily use remote analogies (see Gentner et al., 1997; Holyoak & Thagard 1995). Participants in experiments tend to focus on the sharing of a superficial feature between the source and the target, rather than the relations among features. In his *in vivo* studies of science, Dunbar (1995, 2001, 2002) investigated the ways that scientists use analogies while they are conducting their research and found that scientists use both relational and superficial features when they make analogies. Whether they use superficial or relational features depends on their goals. If their goal is to fix a problem in an experiment, their analogies are based upon superficial features. However, if their goal is to formulate hypotheses, they focus on analogies based upon sets of relations. One important difference between scientists and participants in experiments is that the scientists have deep relational knowledge of the processes that they are investigating and can hence use this relational knowledge to make analogies (see Holyoak, Chapter 13 for a thorough review of analogical reasoning).

Are scientific analogies always useful? Sometimes analogies can lead scientists and students astray. For example, Evelyn Fox-Keller (1985) shows how an analogy between the pulsing of a lighthouse and the activity of the slime mold *dictyostelium* led researchers astray for a number of years. Likewise, the analogy between the solar system (the source) and the structure of the atom (the target) has been shown to be potentially misleading to students taking more advanced courses in physics or chemistry. The solar system analogy has a number of misalignments to the structure of the atom, such as electrons being repelled from each other rather than attracted; moreover, electrons do not have individual orbits like planets but have orbit clouds of electron density. Furthermore, students have serious misconceptions about the nature of the solar system, which can compound their misunderstanding of the nature of the atom (Fischler & Lichtfeld, 1992). While analogy is a powerful tool in science, like all forms of induction, incorrect conclusions can be reached.

Conceptual Change in Science

Scientific knowledge continually accumulates as scientists gather evidence about the natural world.

Over extended time, this knowledge accumulation leads to major revisions, extensions, and new organizational forms for expressing what is known about nature. Indeed, these changes are so substantial that philosophers of science speak of “revolutions” in a variety of scientific domains (Kuhn, 1962). The psychological literature that explores the idea of revolutionary conceptual change can be roughly divided into (a) investigations of how scientists actually make discoveries and integrate those discoveries into existing scientific contexts, and (b) investigations of nonscientists ranging from infants, to children, to students in science classes. In this section we summarize the adult studies of conceptual change, and in the next section we look at its developmental aspects.

Scientific concepts, like all concepts, can be characterized as containing a variety of “knowledge elements”: representations of words, thoughts, actions, objects, and processes. At certain points in the history of science, the accumulated evidence has demanded major shifts in the way these collections of knowledge elements are organized. This “radical conceptual change” process (see Keil, 1999; Nersessian, 1998, 2002; Thagard, 1992; Vosniadou, 1998, for reviews) requires the formation of a new conceptual system that organizes knowledge in new ways, adds new knowledge, and results in a very different conceptual structure. For more recent research on conceptual change, *The International Handbook of Research on Conceptual Change* (Vosniadou, 2008) provides a detailed compendium of theories and controversies within the field.

While conceptual change in science is usually characterized by large-scale changes in concepts that occur over extensive periods of time, it has been possible to observe conceptual change using *in vivo* methodologies. Dunbar (1995) reported a major conceptual shift that occurred in immunologists, where they obtained a series of unexpected findings that forced the scientists to propose a new concept in immunology that in turn forced the change in other concepts. The drive behind this conceptual change was the discovery of a series of different unexpected findings or anomalies that required the scientists to both revise and reorganize their conceptual knowledge. Interestingly, this conceptual change was achieved by a group of scientists reasoning collaboratively, rather than by a scientist working alone. Different scientists tend to work on different aspects of concepts, and also different concepts, that when put together lead to a rapid change in entire conceptual structures.

Overall, accounts of conceptual change in individuals indicate that it is indeed similar to that of conceptual change in entire scientific fields. Individuals need to be confronted with anomalies that their preexisting theories cannot explain before entire conceptual structures are overthrown. However, replacement conceptual structures have to be generated before the old conceptual structure can be discarded. Sometimes, people do not overthrow their original conceptual theories and through their lives maintain their original views of many fundamental scientific concepts. Whether people actively possess naive theories, or whether they appear to have a naive theory because of the demand characteristics of the testing context, is a lively source of debate within the science education community (see Gupta, Hammer, & Redish, 2010).

Scientific Thinking in Children

Well before their first birthday, children appear to know several fundamental facts about the physical world. For example, studies with infants show that they behave as if they understand that solid objects endure over time (e.g., they don't just disappear and reappear, they cannot move through each other, and they move as a result of collisions with other solid objects or the force of gravity (Baillargeon, 2004; Carey, 1985; Cohen & Cashon, 2006; Duschl, Schweingruber, & Shouse, 2007; Gelman & Baillargeon, 1983; Gelman & Kalish, 2006; Mandler, 2004; Metz, 1995; Munakata, Casey, & Diamond, 2004). And even 6-month-olds are able to predict the future location of a moving object that they are attempting to grasp (Von Hofsten, 1980; Von Hofsten, Feng, & Spelke, 2000). In addition, they appear to be able to make nontrivial inferences about causes and their effects (Gopnik et al., 2004).

The similarities between children's thinking and scientists' thinking have an inherent allure and an internal contradiction. The allure resides in the enthusiastic wonder and openness with which both children and scientists approach the world around them. The paradox comes from the fact that different investigators of children's thinking have reached diametrically opposing conclusions about just how “scientific” children's thinking really is. Some claim support for the “child as a scientist” position (Brewer & Samarpungavan, 1991; Gelman & Wellman, 1991; Gopnik, Meltzoff, & Kuhl, 1999; Karmiloff-Smith, 1988; Sodian, Zaitchik, & Carey, 1991; Samarpungavan, 1992), while others offer serious challenges to the view (Fay & Klahr, 1996;

Kern, Mirels, & Hinshaw, 1983; Kuhn, Amsel, & O'Laughlin, 1988; Schauble & Glaser, 1990; Siegler & Liebert, 1975.) Such fundamentally incommensurate conclusions suggest that this very field—children's scientific thinking—is ripe for a conceptual revolution!

A recent comprehensive review (Duschl, Schweingruber, & Shouse, 2007) of what children bring to their science classes offers the following concise summary of the extensive developmental and educational research literature on children's scientific thinking:

- Children entering school already have substantial knowledge of the natural world, much of which is implicit.
- What children are capable of at a particular age is the result of a complex interplay among maturation, experience, and instruction. What is developmentally appropriate is not a simple function of age or grade, but rather is largely contingent on children's prior opportunities to learn.
- Students' knowledge and experience play a critical role in their science learning, influencing four aspects of science understanding, including (*a*) knowing, using, and interpreting scientific explanations of the natural world; (*b*) generating and evaluating scientific evidence and explanations, (*c*) understanding how scientific knowledge is developed in the scientific community, and (*d*) participating in scientific practices and discourse.
- Students learn science by actively engaging in the practices of science.

In the previous section of this article we discussed conceptual change with respect to scientific fields and undergraduate science students. However, the idea that children undergo radical conceptual change in which old "theories" need to be overthrown and reorganized has been a central topic in understanding changes in scientific thinking in both children and across the life span. This radical conceptual change is thought to be necessary for acquiring many new concepts in physics and is regarded as the major source of difficulty for students. The factors that are at the root of this conceptual shift view have been difficult to determine, although there have been a number of studies in cognitive development (Carey, 1985; Chi, 1992; Chi & Roscoe, 2002), in the history of science (Thagard, 1992), and in physics education (Clement, 1982; Mestre, 1991) that give detailed accounts of the changes

in knowledge representation that occur while people switch from one way of representing scientific knowledge to another.

One area where students show great difficulty in understanding scientific concepts is physics. Analyses of students' changing conceptions, using interviews, verbal protocols, and behavioral outcome measures, indicate that large-scale changes in students' concepts occur in physics education (see McDermott & Redish, 1999, for a review of this literature). Following Kuhn (1962), many researchers, but not all, have noted that students' changing conceptions resemble the sequences of conceptual changes in physics that have occurred in the history of science. These notions of radical paradigm shifts and ensuing incompatibility with past knowledge-states have called attention to interesting parallels between the development of particular scientific concepts in children and in the history of physics. Investigations of nonphysicists' understanding of motion indicate that students have extensive misunderstandings of motion. Some researchers have interpreted these findings as an indication that many people hold erroneous beliefs about motion similar to a medieval "impetus" theory (McCloskey, Caramazza, & Green, 1980). Furthermore, students appear to maintain "impetus" notions even after one or two courses in physics. In fact, some authors have noted that students who have taken one or two courses in physics can perform worse on physics problems than naive students (Mestre, 1991). Thus, it is only after extensive learning that we see a conceptual shift from impetus theories of motion to Newtonian scientific theories.

How one's conceptual representation shifts from "naive" to Newtonian is a matter of contention, as some have argued that the shift involves a radical conceptual change, whereas others have argued that the conceptual change is not really complete. For example, Kozhevnikov and Hegarty (2001) argue that much of the naive impetus notions of motion are maintained at the expense of Newtonian principles even with extensive training in physics. However, they argue that such impetus principles are maintained at an implicit level. Thus, although students can give the correct Newtonian answer to problems, their reaction times to respond indicate that they are also using impetus theories when they respond. An alternative view of conceptual change focuses on whether there are real conceptual changes at all. Gupta, Hammer and Redish (2010) and Disessa (2004) have conducted detailed investigations of changes

in physics students' accounts of phenomena covered in elementary physics courses. They have found that rather than students possessing a naive theory that is replaced by the standard theory, many introductory physics students have no stable physical theory but rather construct their explanations from elementary pieces of knowledge of the physical world.

Computational Approaches to Scientific Thinking

Computational approaches have provided a more complete account of the scientific mind. Computational models provide specific detailed accounts of the cognitive processes underlying scientific thinking. Early computational work consisted of taking a scientific discovery and building computational models of the reasoning processes involved in the discovery. Langley, Simon, Bradshaw, and Zytkow (1987) built a series of programs that simulated discoveries such as those of Copernicus, Bacon, and Stahl. These programs had various inductive reasoning algorithms built into them, and when given the data that the scientists used, they were able to propose the same rules. Computational models make it possible to propose detailed models of the cognitive subcomponents of scientific thinking that specify exactly how scientific theories are generated, tested, and amended (see Darden, 1997, and Shrager & Langley, 1990, for accounts of this branch of research). More recently, the incorporation of scientific knowledge into computer programs has resulted in a shift in emphasis from using programs to simulate discoveries to building programs that are used to help scientists make discoveries. A number of these computer programs have made novel discoveries. For example, Valdes-Perez (1994) has built systems for discoveries in chemistry, and Fajtlowicz has done this in mathematics (Erdos, Fajtlowicz, & Staton, 1991).

These advances in the fields of computer discovery have led to new fields, conferences, journals, and even departments that specialize in the development of programs devised to search large databases in the hope of making new scientific discoveries (Langley, 2000, 2002). This process is commonly known as "data mining." This approach has only proved viable relatively recently, due to advances in computer technology. Biswal et al. (2010), Mitchell (2009), and Yang (2009) provide recent reviews of data mining in different scientific fields. Data mining is at the core of drug discovery, our understanding of the human genome, and our understanding

of the universe for a number of reasons. First, vast databases concerning drug actions, biological processes, the genome, the proteome, and the universe itself now exist. Second, the development of high throughput data-mining algorithms makes it possible to search for new drug targets, novel biological mechanisms, and new astronomical phenomena in relatively short periods of time. Research programs that took decades, such as the development of penicillin, can now be done in days (Yang, 2009).

Another recent shift in the use of computers in scientific discovery has been to have both computers and people make discoveries together, rather than expecting that computers make an entire scientific discovery. Now instead of using computers to mimic the entire scientific discovery process as used by humans, computers can use powerful algorithms that search for patterns on large databases and provide the patterns to humans who can then use the output of these computers to make discoveries, ranging from the human genome to the structure of the universe. However, there are some robots such as ADAM, developed by King (2011), that can actually perform the entire scientific process, from the generation of hypotheses, to the conduct of experiments and the interpretation of results, with little human intervention. The ongoing development of scientific robots by some scientists (King et al., 2009) thus continues the tradition started by Herbert Simon in the 1960s. However, many of the controversies as to whether the robot is a "real scientist" or not continue to the present (Evans & Rzhetsky, 2010, Gianfelici, 2010; Haufe, Elliott, Burian, & O'Malley, 2010; O' Malley, 2011).

Scientific Thinking and Science Education

Accounts of the nature of science and research on scientific thinking have had profound effects on science education along many levels, particularly in recent years. Science education from the 1900s until the 1970s was primarily concerned with teaching students both the content of science (such as Newton's laws of motion) or the methods that scientists need to use in their research (such as using experimental and control groups). Beginning in the 1980s, a number of reports (e.g., American Association for the Advancement of Science, 1993; National Commission on Excellence in Education, 1983; Rutherford & Ahlgren, 1991) stressed the need for teaching scientific thinking skills rather than just methods and content. The addition of scientific thinking skills to the science curriculum

from kindergarten through adulthood was a major shift in focus. Many of the particular scientific thinking skills that have been emphasized are skills covered in previous sections of this chapter, such as teaching deductive and inductive thinking strategies. However, rather than focusing on one particular skill, such as induction, researchers in education have focused on how the different components of scientific thinking are put together in science. Furthermore, science educators have focused upon situations where science is conducted collaboratively, rather than being the product of one person thinking alone. These changes in science education parallel changes in methodologies used to investigate science, such as analyzing the ways that scientists think and reason in their laboratories.

By looking at science as a complex multilayered and group activity, many researchers in science education have adopted a constructivist approach. This approach sees learning as an active rather than a passive process, and it suggests that students learn through constructing their scientific knowledge. We will first describe a few examples of the constructivist approach to science education. Following that, we will address several lines of work that challenge some of the assumptions of the constructivist approach to science education.

Often the goal of constructivist science education is to produce conceptual change through guided instruction where the teacher or professor acts as a guide to discovery, rather than the keeper of all the facts. One recent and influential approach to science education is the inquiry-based learning approach. Inquiry-based learning focuses on posing a problem or a puzzling event to students and asking them to propose a hypothesis that could explain the event. Next, the student is asked to collect data that test the hypothesis, make conclusions, and then reflect upon both the original problem and the thought processes that they used to solve the problem. Often students use computers that aid in their construction of new knowledge. The computers allow students to learn many of the different components of scientific thinking. For example, Reiser and his colleagues have developed a learning environment for biology, where students are encouraged to develop hypotheses in groups, codify the hypotheses, and search databases to test these hypotheses (Reiser et al., 2001).

One of the myths of science is the lone scientist suddenly shouting “Eureka, I have made a discovery!” Instead, *in vivo* studies of scientists (e.g.,

Dunbar, 1995, 2002), historical analyses of scientific discoveries (Nersessian, 1999), and studies of children learning science at museums have all pointed to collaborative scientific discovery mechanisms as being one of the driving forces of science (Atkins et al., 2009; Azmitia & Crowley, 2001). What happens during collaborative scientific thinking is that there is usually a triggering event, such as an unexpected result or situation that a student does not understand. This results in other members of the group adding new information to the person’s representation of knowledge, often adding new inductions and deductions that both challenge and transform the reasoner’s old representations of knowledge (Chi & Roscoe, 2002; Dunbar, 1998). Social mechanisms play a key component in fostering changes in concepts that have been ignored in traditional cognitive research but are crucial for both science and science education. In science education there has been a shift to collaborative learning, particularly at the elementary level; however, in university education, the emphasis is still on the individual scientist. As many domains of science now involve collaborations across scientific disciplines, we expect the explicit teaching of heuristics for collaborative science to increase.

What is the best way to teach and learn science? Surprisingly, the answer to this question has been difficult to uncover. For example, toward the end of the last century, influenced by several thinkers who advocated a constructivist approach to learning, ranging from Piaget (Beilin, 1994) to Papert (1980), many schools answered this question by adopting a philosophy dubbed “discovery learning.” Although a clear operational definition of this approach has yet to be articulated, the general idea is that children are expected to learn science by reconstructing the processes of scientific discovery—in a range of areas from computer programming to chemistry to mathematics. The premise is that letting students discover principles on their own, set their own goals, and collaboratively explore the natural world produces deeper knowledge that transfers widely.

The research literature on science education is far from consistent in its use of terminology. However, our reading suggests that “discovery learning” differs from “inquiry-based learning” in that few, if any, guidelines are given to students in discovery learning contexts, whereas in inquiry learning, students are given hypotheses and specific goals to achieve (see the second paragraph of this section for a definition of inquiry-based learning). Even though thousands of schools have adopted discovery learning as an

alternative to more didactic approaches to teaching and learning, the evidence showing that it is more effective than traditional, direct, teacher-controlled instructional approaches is mixed, at best (Lorch et al., 2010; Minner, Levy, & Century, 2010). In several cases where the distinctions between direct instruction and more open-ended constructivist instruction have been clearly articulated, implemented, and assessed, direct instruction has proven to be superior to the alternatives (Chen & Klahr, 1999; Toth, Klahr, & Chen, 2000). For example, in a study of third- and fourth-grade children learning about experimental design, Klahr and Nigam (2004) found that many more children learned from direct instruction than from discovery learning. Furthermore, they found that among the few children who did manage to learn from a discovery method, there was no better performance on a far transfer test of scientific reasoning than that observed for the many children who learned from direct instruction.

The idea of children learning most of their science through a process of self-directed discovery has some romantic appeal, and it may accurately describe the personal experience of a handful of world-class scientists. However, the claim has generated some contentious disagreements (Kirschner, Sweller, & Clark, 2006; Klahr, 2010; Taber, 2009; Tobias & Duffy, 2009), and the jury remains out on the extent to which most children can learn science that way.

Conclusions and Future Directions

The field of scientific thinking is now a thriving area of research with strong underpinnings in cognitive psychology and cognitive science. In recent years, a new professional society has been formed that aims to facilitate this integrative and interdisciplinary approach to the psychology of science, with its own journal and regular professional meetings.¹ Clearly the relations between these different aspects of scientific thinking need to be combined in order to produce a truly comprehensive picture of the scientific mind.

While much is known about certain aspects of scientific thinking, much more remains to be discovered. In particular, there has been little contact between cognitive, neuroscience, social, personality, and motivational accounts of scientific thinking. Research in thinking and reasoning has been expanded to use the methods and theories of cognitive neuroscience (see Morrison & Knowlton, Chapter 6). A similar

approach can be taken in exploring scientific thinking (see Dunbar et al., 2007). There are two main reasons for taking a neuroscience approach to scientific thinking. First, functional neuroimaging allows the researcher to look at the entire human brain, making it possible to see the many different sites that are involved in scientific thinking and gain a more complete understanding of the entire range of mechanisms involved in this type of thought. Second, these brain-imaging approaches allow researchers to address fundamental questions in research on scientific thinking, such as the extent to which ordinary thinking in non-scientific contexts and scientific thinking recruit similar versus disparate neural structures of the brain.

Dunbar (2009) has used some novel methods to explore Simon's assertion, cited at the beginning of this chapter, that scientific thinking uses the same cognitive mechanisms that all human beings possess (rather than being an entirely different type of thinking) but combines them in ways that are specific to a particular aspect of science or a specific discipline of science. For example, Fugelsang and Dunbar (2009) compared causal reasoning when two colliding circular objects were labeled balls or labeled subatomic particles. They obtained different brain activation patterns depending on whether the stimuli were labeled balls or subatomic particles. In another series of experiments, Dunbar and colleagues used functional magnetic resonance imaging (fMRI) to study patterns of activation in the brains of students who have and who have not undergone conceptual change in physics. For example, Fugelsang and Dunbar (2005) and Dunbar et al. (2007) have found differences in the activation of specific brain sites (such as the anterior cingulate) for students when they encounter evidence that is inconsistent with their current conceptual understandings. These initial cognitive neuroscience investigations have the potential to reveal the ways that knowledge is organized in the scientific brain and provide detailed accounts of the nature of the representation of scientific knowledge. Petitto and Dunbar (2004) proposed the term "educational neuroscience" for the integration of research on education, including science education, with research on neuroscience. However, see Fitzpatrick (in press) for a very different perspective on whether neuroscience approaches are relevant to education. Clearly, research on the scientific brain is just beginning. We as scientists are beginning to get a reasonable grasp of the inner workings of the subcomponents of the scientific mind (i.e., problem solving, analogy, induction). However, great advances remain to be made

concerning how these processes interact so that scientific discoveries can be made. Future research will focus on both the collaborative aspects of scientific thinking and the neural underpinnings of the scientific mind.

Note

1. The International Society for the Psychology of Science and Technology (ISPST). Available at <http://www.ispstonline.org/>

References

- Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, 54, 299–352.
- American Association for the Advancement of Science. (1993). *Benchmarks for scientific literacy*. New York: Oxford University Press.
- Atkins, L. J., Velez, L., Goudy, D., & Dunbar, K. N. (2009). The unintended effects of interactive objects and labels in the science museum. *Science Education*, 93, 161–184.
- Azmanita, M. A., & Crowley, K. (2001). The rhythms of scientific thinking: A study of collaboration in an earthquake microworld. In K. Crowley, C. Schunn, & T. Okada (Eds.), *Designing for science: Implications from everyday, classroom, and professional settings* (pp. 45–72). Mahwah, NJ: Erlbaum.
- Bacon, F. (1620/1854). *Novum organum* (B. Monatgue, Trans.). Philadelphia, PA: Parry & McMillan.
- Baillargeon, R. (2004). Infants' reasoning about hidden objects: Evidence for event-general and event-specific expectations (article with peer commentaries and response, listed below). *Developmental Science*, 7, 391–424.
- Baker, L. M., & Dunbar, K. (2000). Experimental design heuristics for scientific discovery: The use of baseline and known controls. *International Journal of Human Computer Studies*, 53, 335–349.
- Beilin, H. (1994). Jean Piaget's enduring contribution to developmental psychology. In R. D. Parke, P. A. Ornstein, J. J. Rieser, & C. Zahn-Waxler (Eds.), *A century of developmental psychology* (pp. 257–290). Washington, DC US: American Psychological Association.
- Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S.M., et al. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences of the United States of America*, 107, 4734–4739.
- Brewer, W. F., & Samarapungavan, A. (1991). Children's theories vs. scientific theories: Differences in reasoning or differences in knowledge? In R. R. Hoffman & D. S. Palermo (Eds.), *Cognition and the symbolic processes: Applied and ecological perspectives* (pp. 209–232). Hillsdale, NJ: Erlbaum.
- Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: NY Science Editions.
- Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.
- Caruthers, P., Stich, S., & Siegal, M. (2002). *The cognitive basis of science*. New York: Cambridge University Press.
- Chi, M. (1992). Conceptual change within and across ontological categories: Examples from learning and discovery in science. In R. Giere (Ed.), *Cognitive models of science* (pp. 129–186). Minneapolis: University of Minnesota Press.
- Chi, M. T. H., & Roscoe, R. D. (2002). The processes and challenges of conceptual change. In M. Limon & L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice* (pp. 3–27). Amsterdam, Netherlands: Kluwer Academic Publishers.
- Chen, Z., & Klahr, D. (1999). All other things being equal: Children's acquisition of the control of variables strategy. *Child Development*, 70(5), 1098–1120.
- Clement, J. (1982). Students' preconceptions in introductory mechanics. *American Journal of Physics*, 50, 66–71.
- Cohen, L. B., & Cashon, C. H. (2006). Infant cognition. In W. Damon & R. M. Lerner (Series Eds.) & D. Kuhn & R. S. Siegler (Vol. Eds.), *Handbook of child psychology. Vol. 2: Cognition, perception, and language* (6th ed., pp. 214–251). New York: Wiley.
- National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: US Department of Education.
- Crick, F. H. C. (1988). *What mad pursuit: A personal view of science*. New York: Basic Books.
- Darden, L. (2002). Strategies for discovering mechanisms: Schema instantiation, modular subassembly, forward chaining/backtracking. *Philosophy of Science*, 69, S354–S365.
- Davenport, J. L., Yaron, D., Klahr, D., & Koedinger, K. (2008). Development of conceptual understanding and problem solving expertise in chemistry. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 751–756). Austin, TX: Cognitive Science Society.
- diSessa, A. A. (2004). Contextuality and coordination in conceptual change. In E. Redish & M. Vicentini (Eds.), *Proceedings of the International School of Physics "Enrico Fermi": Research on physics education* (pp. 137–156). Amsterdam, Netherlands: ISO Press/Italian Physics Society.
- Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg, & J. Davidson (Eds.), *Mechanisms of insight* (pp. 365–395). Cambridge, MA: MIT press.
- Dunbar, K. (1997). How scientists think: Online creativity and conceptual change in science. In T. B. Ward, S. M. Smith, & S. Vaid (Eds.), *Conceptual structures and processes: Emergence, discovery and change* (pp. 461–494). Washington, DC: American Psychological Association.
- Dunbar, K. (1998). Problem solving. In W. Bechtel & G. Graham (Eds.), *A companion to cognitive science* (pp. 289–298). London: Blackwell.
- Dunbar, K. (1999). The scientist *In Vivo*: How scientists think and reason in the laboratory. In L. Magnani, N. Nersessian, & P. Thagard (Eds.), *Model-based reasoning in scientific discovery* (pp. 85–100). New York: Plenum.
- Dunbar, K. (2001). The analogical paradox: Why analogy is so easy in naturalistic settings, yet so difficult in the psychology laboratory. In D. Gentner, K. J. Holyoak, & B. Kokinov (Eds.), *Analogy: Perspectives from cognitive science* (pp. 313–334). Cambridge, MA: MIT press.
- Dunbar, K. (2002). Science as category: Implications of *In Vivo* science for theories of cognitive development, scientific discovery, and the nature of science. In P. Caruthers, S. Stich, & M. Siegel (Eds.), *Cognitive models of science* (pp. 154–170). New York: Cambridge University Press.
- Dunbar, K. (2009). The biology of physics: What the brain reveals about our physical understanding of the world. In M. Sabella, C. Henderson, & C. Singh. (Eds.), *Proceedings of the Physics Education Research Conference* (pp. 15–18). Melville, NY: American Institute of Physics.

- Dunbar, K., & Fugelsang, J. (2004). Causal thinking in science: How scientists and students interpret the unexpected. In M. E. Gorman, A. Kincannon, D. Gooding, & R. D. Tweney (Eds.), *New directions in scientific and technical thinking* (pp. 57–59). Mahwah, NJ: Erlbaum.
- Dunbar, K., Fugelsang, J., & Stein, C. (2007). Do naïve theories ever go away? In M. Lovett & P. Shah (Eds.), *Thinking with Data: 33rd Carnegie Symposium on Cognition* (pp. 193–206). Mahwah, NJ: Erlbaum.
- Dunbar, K., & Sussman, D. (1995). Toward a cognitive account of frontal lobe function: Simulating frontal lobe deficits in normal subjects. *Annals of the New York Academy of Sciences*, 769, 289–304.
- Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (Eds.). (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: National Academies Press.
- Einstein, A. (1950). *Out of my later years*. New York: Philosophical Library.
- Erdos, P., Fajtlowicz, S., & Staton, W. (1991). Degree sequences in the triangle-free graphs, *Discrete Mathematics*, 92(91), 85–88.
- Evans, J., & Rzhetsky, A. (2010). Machine science. *Science*, 329, 399–400.
- Fay, A., & Klahr, D. (1996). Knowing about guessing and guessing about knowing: Preschoolers' understanding of indeterminacy. *Child Development*, 67, 689–716.
- Fischler, H., & Lichtenfeld, M. (1992). Modern physics and students' conceptions. *International Journal of Science Education*, 14, 181–190.
- Fitzpatrick, S. M. (in press). Functional brain imaging: Neuroturn or wrong turn? In M. M. Littlefield & J. M. Johnson (Eds.), *The neuroscientific turn: Transdisciplinarity in the age of the brain*. Ann Arbor: University of Michigan Press.
- Fox-Keller, E. (1985). *Reflections on gender and science*. New Haven, CT: Yale University Press.
- Fugelsang, J., & Dunbar, K. (2005). Brain-based mechanisms underlying complex causal thinking. *Neuropsychologia*, 43, 1204–1213.
- Fugelsang, J., & Dunbar, K. (2009). Brain-based mechanisms underlying causal reasoning. In E. Kraft (Ed.), *Neural correlates of thinking* (pp. 269–279). Berlin, Germany: Springer.
- Fugelsang, J., Stein, C., Green, A., & Dunbar, K. (2004). Theory and data interactions of the scientific mind: Evidence from the molecular and the cognitive laboratory. *Canadian Journal of Experimental Psychology*, 58, 132–141.
- Galilei, G. (1638/1991). *Dialogues concerning two new sciences* (A. de Salvio & H. Crew, Trans.). Amherst, NY: Prometheus Books.
- Galison, P. (2003). *Einstein's clocks, Poincaré's maps: Empires of time*. New York: W. W. Norton.
- Gelman, R., & Baillargeon, R. (1983). A review of Piagetian concepts. In P. H. Mussen (Series Ed.) & J. H. Flavell & E. M. Markman (Vol. Eds.), *Handbook of child psychology* (4th ed., Vol. 3, pp. 167–230). New York: Wiley.
- Gelman, S. A., & Kalish, C. W. (2006). Conceptual development. In D. Kuhn & R. Siegler (Eds.), *Handbook of child psychology. Vol. 2: Cognition, perception and language* (pp. 687–733). New York: Wiley.
- Gelman, S., & Wellman, H. (1991). Insides and essences. *Cognition*, 38, 214–244.
- Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, 34, 752–775.
- Gentner, D., Brem, S., Ferguson, R. W., Markman, A. B., Levidow, B. B., Wolff, P., & Forbus, K. D. (1997). Analogical reasoning and conceptual change: A case study of Johannes Kepler. *The Journal of the Learning Sciences*, 6(1), 3–40.
- Gentner, D., Holyoak, K. J., & Kokinov, B. (2001). *The analogical mind: Perspectives from cognitive science*. Cambridge, MA: MIT Press.
- Gentner, D., & Jeziorski, M. (1993). The shift from metaphor to analogy in western science. In A. Ortony (Ed.), *Metaphor and thought* (2nd ed., pp. 447–480). Cambridge, England: Cambridge University Press.
- Gianfelici, F. (2010). Machine science: Truly machine-aided science. *Science*, 330, 317–319.
- Giere, R. (1993). *Cognitive models of science*. Minneapolis: University of Minnesota Press.
- Gopnik, A. N., Meltzoff, A. N., & Kuhl, P. K. (1999). *The scientist in the crib: Minds, brains and how children learn*. New York: Harper Collins.
- Gorman, M. E. (1989). Error, falsification and scientific inference: An experimental investigation. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 41A, 385–412.
- Gorman, M. E., Kincannon, A., Gooding, D., & Tweney, R. D. (2004). *New directions in scientific and technical thinking*. Mahwah, NJ: Erlbaum.
- Gupta, A., Hammer, D., & Redish, E. F. (2010). The case for dynamic models of learners' ontologies in physics. *Journal of the Learning Sciences*, 19(3), 285–321.
- Haufe, C., Elliott, K. C., Burian, R., & O'Malley, M. A. (2010). Machine science: What's missing. *Science*, 330, 318–320.
- Hecht, E. (2011). On defining mass. *The Physics Teacher*, 49, 40–43.
- Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin and Review*, 7, 569–592.
- Holyoak, K. J., & Thagard, P. (1995). *Mental leaps*. Cambridge, MA: MIT Press.
- Karmiloff-Smith, A. (1988). The child is a theoretician, not an inductivist. *Mind and Language*, 3, 183–195.
- Keil, F. C. (1999). Conceptual change. In R. Wilson & F. Keil (Eds.), *The MIT encyclopedia of cognitive science*. (pp. 179–182) Cambridge, MA: MIT press.
- Kern, L. H., Mirels, H. L., & Hinshaw, V. G. (1983). Scientists' understanding of propositional logic: An experimental investigation. *Social Studies of Science*, 13, 131–146.
- King, R. D. (2011). Rise of the robo scientists. *Scientific American*, 304(1), 73–77.
- King, R. D., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., et al. (2009). The automation of science. *Science*, 324, 85–89.
- Kirschner, P. A., Sweller, J., & Clark, R. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 41, 75–86.
- Klahr, D. (2000). *Exploring science: The cognition and development of discovery processes*. Cambridge, MA: MIT Press.
- Klahr, D. (2010). Coming up for air: But is it oxygen or phlogiston? A response to Taber's review of constructivist instruction: Success or failure? *Education Review*, 13(13), 1–6.

- Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, 12, 1–48.
- Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: effects of direct instruction and discovery learning. *Psychological Science*, 15(10), 661–667.
- Klahr, D. & Masnick, A. M. (2002). Explaining, but not discovering, abduction. Review of L. Magnani (2001) abduction, reason, and science: Processes of discovery and explanation. *Contemporary Psychology*, 47, 740–741.
- Klahr, D., & Simon, H. (1999). Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin*, 125, 524–543.
- Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94, 211–228.
- Kozhevnikov, M., & Hegarty, M. (2001). Impetus beliefs as default heuristic: Dissociation between explicit and implicit knowledge about motion. *Psychonomic Bulletin and Review*, 8, 439–453.
- Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.
- Kuhn, D., Amsel, E., & O'Laughlin, M. (1988). *The development of scientific thinking skills*. Orlando, FL: Academic Press.
- Kulkarni, D., & Simon, H. A. (1988). The processes of scientific discovery: The strategy of experimentation. *Cognitive Science*, 12, 139–176.
- Langley, P. (2000). Computational support of scientific discovery. *International Journal of Human-Computer Studies*, 53, 393–410.
- Langley, P. (2002). Lessons for the computational discovery of scientific knowledge. In *Proceedings of the First International Workshop on Data Mining Lessons Learned* (pp. 9–12).
- Langley, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative processes*. Cambridge, MA: MIT Press.
- Lorch, R. F., Jr., Lorch, E. P., Calderhead, W. J., Dunlap, E. E., Hodell, E. C., & Freer, B. D. (2010). Learning the control of variables strategy in higher and lower achieving classrooms: Contributions of explicit instruction and experimentation. *Journal of Educational Psychology*, 102(1), 90–101.
- Magnani, L., Carnielli, W., & Pizzi, C., (Eds.) (2010). *Model-based reasoning in science and technology: Abduction, logic, and computational discovery*. Series *Studies in Computational Intelligence* (Vol. 314). Heidelberg/Berlin: Springer.
- Mandler, J.M. (2004). *The foundations of mind: Origins of conceptual thought*. Oxford, England: Oxford University Press.
- Macpherson, R., & Stanovich, K. E. (2007). Cognitive ability, thinking dispositions, and instructional set as predictors of critical thinking. *Learning and Individual Differences*, 17, 115–127.
- McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science*, 210, 1139–1141.
- McDermott, L. C., & Redish, L. (1999). Research letter on physics education research. *American Journal of Physics*, 67, 755.
- Mestre, J. P. (1991). Learning and instruction in pre-college physical science. *Physics Today*, 44, 56–62.
- Metz, K. E. (1995). Reassessment of developmental constraints on children's science instruction. *Review of Educational Research*, 65(2), 93–127.
- Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction—what is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching*, 47(4), 474–496.
- Mitchell, T. M. (2009). Mining our reality. *Science*, 326, 1644–1645.
- Mitroff, I. (1974). *The subjective side of science*. Amsterdam, Netherlands: Elsevier.
- Munakata, Y., Casey, B. J., & Diamond, A. (2004). Developmental cognitive neuroscience: Progress and potential. *Trends in Cognitive Sciences*, 8, 122–128.
- Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology*, 29, 89–95.
- Nersessian, N. (1998). Conceptual change. In W. Bechtel, & G. Graham (Eds.), *A companion to cognitive science* (pp. 157–166). London, England: Blackwell.
- Nersessian, N. (1999). Models, mental models, and representations: Model-based reasoning in conceptual change. In L. Magnani, N. Nersessian, & P. Thagard (Eds.), *Model-based reasoning in scientific discovery* (pp. 5–22). New York: Plenum.
- Nersessian, N. J. (2002). The cognitive basis of model-based reasoning in science. In P. Carruthers, S. Stich, & M. Siegal (Eds.), *The cognitive basis of science* (pp. 133–152). New York: Cambridge University Press.
- Nersessian, N. J. (2008). *Creating scientific concepts*. Cambridge, MA: MIT Press.
- O' Malley, M. A. (2011). Exploration, iterativity and kludging in synthetic biology. *Comptes Rendus Chimie*, 14(4), 406–412.
- Papert, S. (1980) Mindstorms: Children computers and powerful ideas. New York: Basic Books.
- Penner, D. E., & Klahr, D. (1996). When to trust the data: Further investigations of system error in a scientific reasoning task. *Memory and Cognition*, 24(5), 655–668.
- Petitto, L. A., & Dunbar, K. (2004). New findings from educational neuroscience on bilingual brains, scientific brains, and the educated mind. In K. Fischer & T. Katzir (Eds.), *Building usable knowledge in mind, brain, and education*. Cambridge, England: Cambridge University Press.
- Popper, K. R. (1959). *The logic of scientific discovery*. London, England: Hutchinson.
- Qin, Y., & Simon, H.A. (1990). Laboratory replication of scientific discovery processes. *Cognitive Science*, 14, 281–312.
- Reiser, B. J., Tabak, I., Sandoval, W. A., Smith, B., Steinmuller, F., & Leone, T. J., (2001). BGuILE: Strategic and conceptual scaffolds for scientific inquiry in biology classrooms. In S. M. Carver & D. Klahr (Eds.), *Cognition and instruction: Twenty-five years of progress* (pp. 263–306). Mahwah, NJ: Erlbaum.
- Riordan, M., Rowson, P. C., & Wu, S. L. (2001). The search for the higgs boson. *Science*, 291, 259–260.
- Rutherford, F.J., & Ahlgren, A. (1991). *Science for all Americans*. New York: Oxford University Press.
- Samarapungavan, A. (1992). Children's judgments in theory choice tasks: Scientific rationality in childhood. *Cognition*, 45, 1–32.
- Schauble, L., & Glaser, R. (1990). Scientific thinking in children and adults. In D. Kuhn (Ed.), *Developmental perspectives on teaching and learning thinking skills*. Contributions to

- Human Development*, (Vol. 21, pp. 9–26). Basel, Switzerland: Karger.
- Schunn, C. D., & Klahr, D. (1995). A 4-space model of scientific discovery. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp. 106–111). Mahwah, NJ: Erlbaum.
- Schunn, C. D., & Klahr, D. (1996). The problem of problem spaces: When and how to go beyond a 2-space model of scientific discovery. Part of symposium on Building a theory of problem solving and scientific discovery: How big is N in N-space search? In *Proceedings of the 18th Annual Conference of the Cognitive Science Society* (pp. 25–26). Mahwah, NJ: Erlbaum.
- Shrager, J., & Langley, P. (1990). *Computational models of scientific discovery and theory formation*. San Mateo, CA: Morgan Kaufmann.
- Siegler, R. S., & Liebert, R. M. (1975). Acquisition of formal scientific reasoning by 10- and 13-year-olds: Designing a factorial experiment. *Developmental Psychology, 11*, 401–412.
- Simon, H. A. (1977). *Models of discovery*. Dordrecht, Netherlands: D. Reidel Publishing.
- Simon, H. A., Langley, P., & Bradshaw, G. L. (1981). Scientific discovery as problem solving. *Synthese, 47*, 1–27.
- Simon, H. A., & Lea, G. (1974). Problem solving and rule induction. In H. Simon (Ed.), *Models of thought* (pp. 329–346). New Haven, CT: Yale University Press.
- Smith, E. E., Shafir, E., & Osherson, D. (1993). Similarity, plausibility, and judgments of probability. *Cognition. Special Issue: Reasoning and decision making, 49*, 67–96.
- Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development, 62*, 753–766.
- Taber, K. S. (2009). Constructivism and the crisis in U.S. science education: An essay review. *Education Review, 12*(12), 1–26.
- Thagard, P. (1992). *Conceptual revolutions*. Cambridge, MA: MIT Press.
- Thagard, P. (1999). *How scientists explain disease*. Princeton, NJ: Princeton University Press.
- Thagard, P., & Croft, D. (1999). Scientific discovery and technological innovation: Ulcers, dinosaur extinction, and the programming language Java. In L. Magnani, N. Nersessian, & P. Thagard (Eds.), *Model-based reasoning in scientific discovery* (pp. 125–138). New York: Plenum.
- Tobias, S., & Duffy, T. M. (Eds.). (2009). *Constructivist instruction: Success or failure?* New York: Routledge.
- Toth, E. E., Klahr, D., & Chen, Z. (2000) Bridging research and practice: A cognitively-based classroom intervention for teaching experimentation skills to elementary school children. *Cognition and Instruction, 18*(4), 423–459.
- Tweney, R. D. (1989). A framework for the cognitive psychology of science. In B. Gholson, A. Houts, R. A. Neimeyer, & W. Shadish (Eds.), *Psychology of science: Contributions to metascience* (pp. 342–366). Cambridge, England: Cambridge University Press.
- Tweney, R. D., Doherty, M. E., & Mynatt, C. R. (1981). *On scientific thinking*. New York: Columbia University Press.
- Valdes-Perez, R. E. (1994). Conjecturing hidden entities via simplicity and conservation laws: Machine discovery in chemistry. *Artificial Intelligence, 65*(2), 247–280.
- Von Hofsten, C. (1980). Predictive reaching for moving objects by human infants. *Journal of Experimental Child Psychology, 30*, 369–382.
- Von Hofsten, C., Feng, Q., & Spelke, E. S. (2000). Object representation and predictive action in infancy. *Developmental Science, 3*, 193–205.
- Vosniadou, S. (Ed.). (2008). *International handbook of research on conceptual change*. New York: Taylor & Francis.
- Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology, 24*, 535–585.
- Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*, 273–281.
- Wertheimer, M. (1945). *Productive thinking*. New York: Harper.
- Yang, Y. (2009). Target discovery from data mining approaches. *Drug Discovery Today, 14*(3–4), 147–154.

Barbara A. Spellman and Frederick Schauer

Abstract

The legal profession has long claimed that there are process-based differences between legal reasoning—that is, the thinking and reasoning of lawyers and judges—and the reasoning of those without legal training. Whether those claims are sound, however, is a subject of considerable debate. We describe the importance in the legal system of using categorization and analogy, following rules and authority, and the odd task of “fact finding.” We frame these topics within the debate between two views of legal reasoning: the traditional view—that when deciding a case, judges are doing something systematic and logical that only legally trained minds can do; and the Legal Realist view—that judges reason in much the same way as ordinary people do, and that they first come to conclusions and then go back to justify them with the law rather than using the law to produce their conclusions in the first place.

Key Words: legal reasoning, psychology and law, reasoning, analogy, precedent, jury decision making

Introduction

In the 1973 film *The Paper Chase*, the iconic Professor Kingsfield announced to his class of first-year law students: “You teach yourself the law. I train your minds. You come in here with a skull full of mush, and if you survive, you’ll leave thinking like a lawyer.” In claiming to teach students to think like lawyers, Kingsfield echoed the assumptions of centuries of legal ideology. In the 17th century, the great English judge Edward Coke glorified the “artificial reason” of the law (Coke, 1628, ¶ 97b), and from then until now lawyers and judges have believed that legal thinking and reasoning is different from ordinary thinking and reasoning, even from very good ordinary thinking and reasoning. Moreover, the difference, as Kingsfield emphasized, has long been thought to be one of process and not simply of content. It is not only that those with legal training know legal rules that laypeople do not. Rather, lawyers and judges are believed, at least by lawyers and judges, to employ techniques

of argument, reasoning, and decision making that diverge from those of even expert nonlawyer reasoners and decision makers.

Our chapter begins by describing three important distinctions: between what people typically mean by “legal reasoning” and other types of reasoning that occur in the legal system; between two competing views of how such reasoning is done; and between law and fact. The heart of the chapter deals with four thinking and reasoning processes that are common in legal reasoning: following rules, categorization, analogy, and fact finding. We then discuss whether legal decision making requires particular expertise and examine some of the peculiarities of legal decision-making procedures generally. We end with some ideas for future research.

The Who, How, and What of Legal Reasoning

What is meant by “legal reasoning”? Who does it, how is it done, and which parts of it do we think

are unique? We sketch answers to these questions in the text that follows.

“Legal Reasoning” Versus Reasoning Within the Legal System

Legal reasoning, strictly speaking, must be distinguished from the full universe of reasoning and decision making that happens to take place within the legal system. Juries, for example, make decisions in court that have legal consequences, but no one claims that the reasoning of a juror is other than that of the ordinary person, even though the information that jurors receive is structured by legal rules and determinative of legal outcomes. There has been extensive psychological research on jury decision making (for example, Diamond & Rose, 2005; Hastie, 1993), and we discuss some of it in this chapter in the section on “Fact Finding.” But when Coke and Kingsfield were glorifying legal reasoning, they were thinking of lawyers and judges and not of lay jurors. Similarly, police officers, probation officers, and even the legislators who make the laws are undeniably part of the legal system, yet the typical claims about the distinctiveness of legal reasoning do not apply to them. Clearly, the institutions and procedures of the legal system affect decision making, but the traditional claims for the distinctiveness of legal reasoning go well beyond claims of mere institutional and procedural differentiation. The traditional claim is that certain legal professionals—lawyers and judges—genuinely reason differently, rather than employ standard reasoning under different institutional procedures.

Thus, the term “legal reasoning” refers to reasoning by a subset of people involved in the legal system; it also refers to a subset of what that subset of people reason about. Television portrayals notwithstanding, a large part of what lawyers do consists of tasks such as negotiating, drafting contracts, writing wills, and managing uncontested dealings with the administrative bureaucracy. These lawyers’ functions are important in understanding the legal system in its entirety, yet they are rarely alleged to involve distinctive methods of thought, except insofar as they are performed with an eye toward potential legal challenges and litigation. Therefore, we focus in this chapter on trials and appeals because that is the domain about which claims for the distinctiveness of legal reasoning are most prominent.

Two Views of Legal Reasoning

In this chapter we examine forms of reasoning that are allegedly concentrated in, even if not

exclusive to, the legal system. But we also address the long history of skeptical challenges to the legal profession’s traditional claims about the distinctiveness of its methods. From the 1930s to the present, theorists and practitioners typically described as Legal Realists (or just “Realists”) have challenged the belief that legal rules and court precedents substantially influence legal outcomes (Frank, 1930; Llewellyn, 1930; Schlegel, 1980). Rather, say the Realists, legal outcomes are primarily determined by factors other than those that are part of the formal law. These nonlegal factors might include the personality of the judge, for example, as well as the judge’s moral and political ideology and her reactions to the facts of the particular situation presented.

The Realists’ claim that such nonlegal considerations are an important part of judicial decision making should come as little surprise to most psychologists. After all, the Realist challenge is largely consistent with the research on motivated reasoning (Braman, 2010). Because decision makers are often focused on reaching specific desired conclusions, the motivation to reach an antecedently desired conclusion will affect their information search and recall, as well as other components of the decision-making process (Kunda, 1987, 1990; Molden & Higgins, 2005; Chapter 20). Insofar as this research is applicable to judges, then, the Realists would claim that judges are frequently motivated to reach specific outcomes in specific cases for reasons other than the existence of a relevant legal rule. They might, for example, sympathize with one party in the particular case. Or they might believe, more generally, for example, that labor unions should ordinarily prevail against corporations (Kennedy, 1986), or that the police should be supported in their fight against typically guilty defendants, or that commerce flows more smoothly if the norms of the business community rather than the norms of the law are applied to commercial transactions (Twining, 1973). These nonlegal and outcome-focused motivations, say the Realists, would lead judges to retrieve legal rules and precedents selectively in light of that motivation, locating and using only or disproportionately the rules and precedents supporting the result generated by their nonlegal outcome preferences in a particular dispute.

Indeed, the same point is supported by the research on confirmation bias (see Nickerson, 1998, for a review). This research teaches us that both novice and expert decision makers are inclined to design their tasks in ways that yield results consistent with

their initial beliefs (Fiedler, 2011). In light of what we know about motivated reasoning and confirmation bias, therefore, it is plausible that judges often consult the formal law only after having tentatively decided how the case, all or many things other than the law considered, ought to come out. The judges would then select or interpret the formal law to support outcomes reached on other grounds, as the Realists contend, rather than using the formal law to produce those outcomes in the first place, as the traditional view of legal reasoning maintains.

The traditional view of “thinking like a lawyer” does not deny that motivated reasoning and confirmation bias influence the decisions of ordinary people. It does deny, however, that these phenomena are as applicable to expert legal reasoners as they are to laypeople. Indeed, it is telling that nominees for judicial appointments, especially nominees to the Supreme Court testifying before the Senate Judiciary Committee at their confirmation hearings, persistently pretend not to be Realists. They deny that any policy or outcome preferences they might happen to have will influence their judicial votes, claiming instead that their job is simply to follow the law.¹ The judicial nominees thus join the claims of Kingsfield and countless others that the forms of thinking and reasoning that characterize human beings in general are exactly the forms of thinking and reasoning that lawyers and judges are trained to avoid. Whether such avoidance can actually be taught or actually occurs, however, are empirical questions, and not the articles of faith they were for Kingsfield. The question of whether lawyers and judges really are better than laypeople at avoiding the consequences of motivated reasoning, confirmation bias, and other impediments to law-generated results is one that lies at the heart of the traditional claims for the distinctiveness of legal reasoning. In this chapter, we consequently discuss not only the traditional view of legal reasoning but also the research examining the extent to which the model of reasoning described by the traditional view accurately characterizes the arguments of lawyers and the decision making of the judges to whom they argue.

The Distinction Between Law and Fact

The distinction between questions of law and questions of fact is crucial to understanding legal decision making. Indeed, questions of fact are primary in important ways, because the initial question in any legal dispute is the question of what

happened—the question of fact. How fast was the Buick going when it collided with the Toyota? Who came into the bank with a gun and demanded money from the teller? Did the shopkeeper actually promise the customer that the lawnmower she bought would last for 5 years? In typical usage we think of “facts” as things that are known to be true. But in the courtroom, relevant facts may be unknown or in dispute. Thus, the first thing that the “trier of fact,” be it jury or judge, must do is “fact finding”—that is, deciding what actually happened.

Knowing what happened is important and preliminary, but knowing what happened does not answer the *legal* question—the question of what consequences flow from what happened. If a prospective employee proves that the company that did not hire him refuses to hire anyone over the age of 50, has the company violated the law, and, if so, what is the penalty? If the defendant in a murder case drove a getaway car but did not shoot anyone, is he subject to the same criminal penalty as an accomplice who actually did the shooting? If the Buick that someone purchased from a Buick dealer turns out to be defective, is the dealer responsible or only the manufacturer?

In the United States it is common to think of juries as determining questions of fact and judges as deciding questions of law. However, this simple dichotomy is misleading. Although juries generally do not decide questions of law (though they are required to apply the law to the facts in order to reach a verdict), judges *do* decide questions of fact. In many countries, there are no juries at all. And even though most countries with a common-law (English) legal heritage have juries for many criminal trials, only in the United States are there still juries for civil lawsuits between private parties.

Even in the United States, juries are far less common than one would suspect from television portrayals of the legal system. Partly because of settlement and plea bargaining, partly because only certain types of cases involve the right to a jury, partly because sometimes the opposing parties agree to have a judge decide the case, partly because of alternative dispute resolution, and partly because many cases are dismissed or otherwise resolved by judges on legal grounds before trial, jury trials are rare. In fact, only about 1% of initiated cases in the United States reach trial at all, and many of those are tried by a judge sitting without a jury (Galanter, 2004). Often, therefore, the issues of fact as well as law are decided by the judge.

And even when there is a jury, many preliminary factual issues will have been decided by the judge. In criminal cases, for example, factual questions about arguably illegal searches and seizures or confessions—Did the police have probable cause to conduct a search? Was the defendant given the requisite warnings before being interrogated?—are determined by the judge. In civil cases with a jury, judges decide many issues of fact in determining preliminary procedural issues and making rulings on the admissibility of evidence—Did the defendant answer the complaint within the required 20-day period? Can an expert in automobile design testify as an expert about tire failure?

The psychological issues implicated by decisions about disputed questions of fact are not necessarily the same as those involved in determining what the law is. And thus we deal separately, much later, with the psychology of factual determination in law. For now, however, it is worth noting that many of the claims about a distinctively legal reasoning pertain to the resolution of uncertain questions about the law rather than about what happened. Determining what the law requires, especially when the law is uncertain, involves the kind of legal reasoning that Kingsfield celebrated and Supreme Court nominees endorse. Learning how to make such determinations is a large part of the training of lawyers, and a substantial component of legal practice, especially in appellate courts. It is precisely when rules or precedents are unclear or generate uncomfortable outcomes that the use of rules, precedents, analogies, and authority becomes most important, and these are the forms of reasoning that are central to the alleged distinctiveness of legal reasoning. We turn to those forms of reasoning now.

Rules

Following, applying, and interpreting formal, written, and authoritative rules, as well as arguing within a framework of such rules, are important tasks for lawyers and judges, and they are consequently emphasized in the standard picture of legal reasoning. The psychology literature does not address this kind of rule following *per se*; however, to a psychologist the processes involved in deciding “easy” cases seem to involve deductive reasoning (see Evans, Chapter 8), whereas those for “hard” cases seem to involve categorization (see Rips et al., Chapter 11) and analogy (see Holyoak, Chapter 13).

The distinction between easy cases and hard cases is widely discussed in the legal literature. In an easy

case, a single and plainly applicable rule gives unambiguous guidance and, as applied to the situation at hand, appears to give the right result. Suppose a law says: “If someone does A, then he gets consequence B.” Richard comes along, blatantly does A, and then gets consequence B. Rule followed; justice done; everyone (except Richard) is happy. But what we illustrate next is that not all rules are so simple, nor can they be so simply and rewardingly applied. Several types of difficulties can arise, making the application of the rules uncertain or, perhaps, undesirable. What are those difficulties that create hard cases and how do judges resolve them? It depends whether you ask a traditionalist or a Realist.

Defining Hard Cases

There are three kinds of hard cases: ones in which the language of an applicable rule is unclear; ones in which it is unclear which of several rules apply; and ones in which the language of a plainly applicable rule is clear but produces what the interpreter, applier, decision maker, or enforcer of the rule believes is the wrong outcome.

UNCLEAR RULES

Legal rules often do not give a clear answer. A famous example in the legal literature involves a hypothetical rule prohibiting vehicles in a public park (Hart, 1958; Schauer, 2008a). When the question is whether that rule prohibits ordinary cars and trucks, the application of the rule is straightforward. Cars are widely understood to be vehicles, vehicles are prohibited according to the rule, and therefore cars, including this car, are prohibited. That people can and sometimes do reason in such a deductive or syllogistic way when they are given clear rules and presented with clear instances of application is well established (Evans, Barston, & Pollard, 1983; Rips, 2001).

But what about bicycles, baby carriages, wheelchairs, and skateboards, none of which are either clearly vehicles or clearly not vehicles? Faced with such an instance, what would a judge do? One standard view is that the judge would then have discretion to decide the issue as she thought best. Perhaps the judge would try to determine the purpose behind the rule, or perhaps she would try to imagine what the original maker of the rule would have thought should be done in such a case. But whatever the exact nature of the inquiry, the basic idea is that the judge would struggle to determine what the unclear rule *really* means in this situation and would then decide the case accordingly.

The view that judges are searching for guidance from even unclear rules is part of the standard ideology of the lawyers and judges. But that view may be at odds with psychological reality. Freed from the strong constraints of a plainly applicable rule, the research on motivated reasoning suggests that the judge would be likely to decide how, on the basis of a wide range of political, ideological, personal, and contextual factors, she believes the case ought to come out (Braman, 2010). Having come to that conclusion, a conclusion not substantially dependent on the legal rule at all, the judge would then describe that result as being the one most consistent with the purpose behind the rule. And if the judge then looked for evidence of the purpose behind the rule, or evidence of what the rule maker intended in making the rule, much of what we know about confirmation bias (Nickerson, 1998) would suggest that the judge would not engage in the search for purpose or intent with an entirely open mind, but rather would be likely to find the evidence of purpose or intent that supported the outcome the judge had initially preferred.

The latter and more skeptical explanation is entirely consistent with the Legal Realist view about rules. In 1929 Joseph Hutcheson, a Texas-based federal judge, wrote an influential article (Hutcheson, 1929) challenging the traditional picture of legal reasoning. He claimed that it is a mistake to suppose that in the typical case that winds up in court the judge would first look to the text of the rule, the purpose behind the rule, the evidence of legislative intent, and the like in order to decide the case. Rather, Hutcheson argued, the judge would initially, and based largely on the particular facts of the case rather than the law, come up with an initial “hunch” about how the case ought to be decided. Then, and only then, would the judge seek to find a rule to support that result or seek to interpret a fuzzy rule in such a way as to justify that result. Subsequent Realists (e.g., Frank, 1930) reinforced this theme, albeit rarely with systematic empirical research.

Thus, the debate between traditional and Legal Realist view about rule following might also be cast in the language of the contemporary research on dual-process methods of thinking (Evans, 2003, 2008; Sloman, 1996; see Evans, Chapter 8; Stanovich, Chapter 22).² System 1 reasoning is quick and intuitive, whereas System 2 reasoning is more logical, systematic, and deliberative (Stanovich, 1999), and the traditional view of legal reasoning relies heavily on a System 2 model of decision making. The Realist

perspective, as exemplified by Hutcheson’s reference to a “hunch,” sees even judicial reasoning as having heavy doses of quick, intuitive, and perhaps heuristic System 1 decision making. (These are sometimes viewed as two separate reasoning systems and sometimes as the ends of a reasoning continuum.) The question remains as to which method of decision making more accurately reflects the reality of judging. Several legal scholars have suggested that judges, just like ordinary people, often come quickly to an intuitive decision but then sometimes override that decision with deliberation. They state, “The intuitive system appears to have a powerful effect on judges’ decision making” (Guthrie, Rachlinski, & Wistrich, 2007, p. 43) and then suggest various ways in which the legal system should increase the likelihood that judges will use System 2 reasoning in deciding cases. Note, however, that when the systems are in opposition it is not always the case that the intuitive system is wrong and the deliberative system is right; it can also turn out the other way (Evans, 2008).

WHEN RULES PROLIFERATE

The second type of hard cases consists of those to which multiple but inconsistent clear rules apply. Is a truck excluded from the park by the “no vehicles in the park” rule, or is it permitted by another rule authorizing trucks to make deliveries wherever necessary? Such instances of multiple and inconsistent rules make the Realist challenge to the conventional picture especially compelling in a legal system in which many rules might plausibly apply to one event. In countries with civil law systems,³ legislatures attempt to enact explicit and clear legal rules covering all conceivable situations and disputes. Such rules are collected in a comprehensive code; therefore, the existence of multiple and inconsistent rules applying to the same event is, at least in theory, rare. Even the outcome-motivated judge might well find that the law plainly did not support the preferred outcome, and that would almost certainly be the end of the matter.

The situation is different in English-origin common-law countries,⁴ where much of the law is made by judges in the process of deciding particular cases. Law making in common-law systems is less systematic than in civil law countries, and common-law judges and legislatures are less concerned than their civil law counterparts with ensuring that new rules fit neatly with all of the existing legal rules. As a result, it is especially in common-law countries that

multiple and inconsistent rules may apply to the same event, allowing for more decisions that seem to be based on motivated reasoning. Moreover, even when a judge does not have a preferred outcome, when there are multiple potentially applicable rules, the judge's background and training, among other things, will influence which rules are retrieved and which are ignored (Spellman, 2010). In addition, if judges, like other people, seek coherence and consistency in their thinking, they may select legal rules and sources that are consistent with the others they have retrieved and ignore those that would make coherence more difficult (Holyoak & Simon, 1999; Simon, Pham, Le, & Holyoak, 2001).

WHEN RULES GIVE THE WRONG ANSWER

Although there can be problems with vague rules and multiple rules, as described earlier, typically the words of a plainly applicable rule, conventionally interpreted, do indicate an outcome, just as the "no vehicles in the park" rule indicates an outcome in a case involving a standard car or truck. But leading to an obvious outcome is not the end of the story. Because rules are generalizations drafted in advance of specific applications (Schauer, 1991), there is the possibility, as with any generalization, that the rule, if strictly or literally followed, will produce what appears to be a bad result in a specific situation. In *Riggs v. Palmer* (1889), for example, a case decided by the New York Court of Appeals, the pertinent statute provided clearly, and without relevant exception, that anyone named in a will could claim his inheritance upon the death of the testator (i.e., the person who wrote the will). The problem in *Riggs*, however, was that the testator died because his grandson, the beneficiary, had poisoned him, and did so precisely and intentionally in order to claim his inheritance as soon as possible. Thus, the question in *Riggs* was whether a beneficiary who murdered the testator could inherit from him. More generally, the question was whether the justice or equity or fairness of the situation should prevail over the literal wording of the rule.

Cases like *Riggs* are legion, and the issues they present raise important issues about the nature of law and legal decisions (Dworkin, 1986). But they also implicate equally important psychological questions. When a rule points in one direction and the all-things-considered right answer points in another, under what conditions, and how often, will people—be they legally trained or not—put aside their best moral or pragmatic judgment in favor of what the rule commands?

If the traditional story is sound, we would expect those with legal training to attach greater value to the very fact of the existence of a legal rule, and thus to prefer the legally generated but morally or pragmatically wrong result more often than those without such training. It turns out, however, that very little research has addressed precisely this question. On the one hand, research has found that law students (Furgeson, Babcock, & Shane, 2008b) and federal law clerks (recent law school graduates) working for federal judges (Furgeson, Babcock, & Shane, 2008a) are affected by their policy preferences in drawing conclusions about the law. On the other hand, there are data indicating that judges are better able to put aside their ideologies than law students in evaluating evidence (Redding & Reppucci, 1999). Most relevantly, legally trained experimental subjects tend to prefer formal rules of justice more often than those without legal training (Schweitzer et al., 2008). Still, the research can best be described as limited, presumably owing to the difficulties in securing judges and lawyers as experimental subjects. And, of course, any study finding differences between groups along the law-training continuum (laypeople, law students, law clerks, lawyers, judges) must consider not only legal training and experience but also selection and self-selection effects (e.g., who chooses to go into law; who is chosen to become a judge) when drawing causal conclusions.

Deciding Hard Cases

It is important to understand the types of difficulties generated by hard cases because litigation, and especially litigation at the appellate stage, is disproportionately about hard cases. Easy cases are plentiful, at least if we understand "cases" to refer to all disputes or even all instances of application of the law (Schauer, 1985). But if the law is clear and if the clear law produces a plausible or palatable outcome, few people would take the case to court in the first place. Only where two opposing parties each believe they have a reasonable chance of winning will the dispute actually arrive in court, and also, to an even greater extent, when disputants decide whether to appeal. As a result of this legal selection effect (Lederman, 1999; Priest & Klein, 1984), the disputes that produce litigation and judicial opinions will disproportionately represent hard cases, with the easy cases—the straightforward application of clear law—not arriving in court at all.

This selection effect is greatest with respect to decisions by the Supreme Court of the United States,

which can choose the cases it will hear. It is asked to formally decide about 9,000 cases per year but considers only about 70 per year with full written and oral arguments. And with respect to these 70 cases, the existing research, mostly by empirical political scientists, supports the conclusion that the political attitudes of the Justices—how they feel about abortion and affirmative action, for example, as a policy matter—is a far better predictor of how they will vote than is the formal law (Segal & Spaeth, 2004). This research is not experimental; rather, it involves coding Justices on a variety of attributes and coding cases on a variety of attributes and then analyzing what predicts what. For example, Justices are coded on such things as age (at the time of the decision), gender, race, residence, political party at the time of nomination; cases are coded on such things as topic, types of litigants, and the applicability of various precedents and legal rules. The conclusion of much of this research is that we can better predict legal outcomes, at least in the Supreme Court and to some extent in other appellate courts, if we know a judge's prelegal policy preferences than if we understand the applicable rules and precedents. To the extent that this research is sound, therefore, it may support the view that the Supreme Court, ironically to some, is the last place we should look to find distinctively legal reasoning (but see Shapiro, 2009, for a critique of these analyses).

Categorization

Questions about rule following obviously implicate important issues of categorization. Do we categorize a skateboard as a vehicle or as a toy? Do we categorize Elmer Palmer, the young man who murdered his grandfather in order to accelerate his inheritance, as a murderer, as a beneficiary, or possibly even as both?

Because legal outcomes are determined by something preexisting called "the law," those outcomes require placing any new event within an existing category. When the category is specified by a written rule with a clear semantic meaning for the pertinent application, as with the category "vehicle" when applied to standard automobiles in the "no vehicles in the park" rule, the freedom of the decision maker is limited by the plausible extensions of the specified category. Often, however, there is no such clear written rule that is literally applicable to the case at hand, sometimes because the rule is vague (consider the Constitution's requirement that states grant "equal protection of the laws" and the constitutional

prohibition on "cruel and unusual punishments"), sometimes because a case arises within the vague penumbra of a rule (as with the skateboard case under the "no vehicles in the park" rule), and often because in common-law systems the relevant law is not contained in a rule with a fixed verbal formulation but instead is in the body of previous judicial decisions. In such cases the task of categorization is more open ended, and decision makers must make less constrained judgments of similarity and difference in order to determine which existing legal category best fits with a new instance.

Legal Categories

The view that legal reasoning and legal expertise is a matter of using and understanding the categories of the law rather than the categories of the prelegal world is one whose iconic expression comes from an apocryphal anecdote created by Oliver Wendell Holmes:

There is a story of a Vermont justice of the peace before whom a suit was brought by one farmer against another for breaking a churn. The justice took time to consider, and then said that he had looked through the statutes and could find nothing about churning, and gave judgment for the defendant. (Holmes, 1897, pp. 474–475)

The point of the anecdote derives from the fact that a justice of the peace would have been a lay decider of minor controversies, not a real judge with legal training and legal expertise. And thus Holmes can be understood as claiming that only an untrained bumpkin could have imagined that "churn" was the relevant legal category. That this is Holmes's point is made clear shortly thereafter, when he says that

[a]pplications of rudimentary rules of contract or tort are tucked away under the heads of Railroads or Telegraphs or... Shipping..., or are gathered under an arbitrary title which is thought likely to appeal to the practical mind, such as Mercantile Law. If a man goes into law it pays to be a master of it, and to be a master of it means to look straight through all the dramatic incidents and to discern the true basis for prophecy. (Holmes, 1897, p. 475)

For Holmes, "railroad" and "telegraph" are lay categories, and "contract" and "tort" are legal categories, and one mark of legal expertise and legal reasoning is the ability to use legal rather than lay categories. This is still not a very strong claim about

the distinctiveness of legal reasoning, for the difference that Holmes identifies is one of content and not of process. The lawyer does not think or reason differently from the layman, Holmes might be understood as saying, but thinks and reasons the same way, albeit with different categories and thus with different content (Spellman, 2010). If legal reasoning does not involve substantially different processes from ordinary reasoning, the strongest claims of the traditional view of legal reasoning are weakened. But if legal reasoning employs the distinctive categories and content of the law, and if these categories in fact determine many legal outcomes, the strongest claims of Legal Realism are weakened as well. By applying substantially (even if not completely) ordinary reasoning to substantially (even if not completely) law-created content and categories, legal reasoning may turn out to have its own special characteristics, but not in ways that either the traditionalists or the Realists maintained.

Relational Categories

We believe that the categories that the law uses tend to be relational categories—categories created on the basis of the relations that one item has with another, rather than on the basis of the attributes of single items taken in isolation (i.e., involving predicates that take at least two objects). That law is principally concerned with the way in which one person or thing is connected or related to another should not be surprising. After all, the law is about regulating interactions and exchanges among people—that is, relations. Take the category of “contract.” Suppose someone wants to know whether Judy and Jerry have entered into a contract. Nearly all personal details about Judy and Jerry are irrelevant, as are nearly all details about what they have contracted for. What *is* relevant is whether Jerry owned the property, whether Judy made what the law defines as an offer, and whether Jerry responded with what the law defines as acceptance. Similarly, suppose that Beth has done something to Brian. Whether that “something” is being hit, libeled, or kidnapped, it is again typically the relation of what one did to the other that matters. And so too with the questions involved in a finding of negligence: Did John harm James? Did John have a duty of care toward James? Again, relations are key. Note that sometimes it does matter whether the person is under 18 (and so can’t sign a contract) or over 35 (and so is eligible to be President of the United States). And sometimes it matters whether a person

is male or female, Black or White, famous or non-famous. But most of the time it is only the relations between the parties that matter. Indeed, the traditional contrast between the “rule of law” and the “rule of men” [sic] stresses the impersonality of the law, and thus its emphasis on the relational “what” rather than the personal “who.”

Despite the large psychology literature on categorization, there has been relatively little work on relational categories (see Gentner & Kurtz, 2005). However, we do know that just like category members from standard categories prime other category members, category members from relational categories prime other category members (e.g., “bird-nest” primes “bear-cave” by activating the relation “lives-in;” Spellman, Holyoak, & Morrison, 2001). We also know that relations are generally more important than attributes for analogical reasoning. Thus, when someone is trained on which relations exist and matter, analogical reminding can be useful for retrieving analogies that can help make a legal argument. How analogy is used in legal reasoning is the topic of the next section.

Precedent and Analogy

In common-law systems much of the law is not to be found in the explicitly written rules enacted by legislatures or adopted by administrative agencies, but in the decisions of judges. And because when judges reach decisions and thus make law they are expected to take account of previous decisions—precedents—the interpretation of precedents is an important part of common-law decision making. In common-law systems, and increasingly in civil law systems, law develops incrementally as decisions in particular cases build on previous decisions. Understanding how to use previous decisions to make an argument or decision in the current dispute is consequently a substantial component of legal reasoning. Previous decisions play a large role in legal reasoning, but they do so in two very different ways (Schauer, 2008c).

“Vertical” Precedent

First, and possibly of less significance in hard cases, is the obligation of a judge to follow the decision of a higher controlling court (hence “vertical”) even if she disagrees with that decision. This is the strong form of constraint by precedent, and it resembles the constraints of an explicitly written rule. When there is a previous decision on the same question (just as when there is an explicit rule plainly covering some

application), the law *tells* the judge what her decision should be. Consider, for example, the obligations of judges with respect to the Supreme Court's decision in *Miranda v. Arizona* (1966), the case in which the Court required police officers to advise a suspect in custody of his rights to remain silent and have a lawyer prior to questioning. *Miranda* was controversial when it was decided and has remained controversial since. Many citizens, police officers, and even judges believe that *Miranda* was a mistaken decision. Nevertheless, a judge in a court below the Supreme Court is not permitted to substitute her judgment for that of the higher court. If the question arises in a lower court as to whether the statements of a suspect who was not advised of his rights can be used against him, the lower court judge who thinks that the answer to this question ought to be "yes" is obliged by the Supreme Court's decision in *Miranda* to answer "no." Obviously there will be difficult cases in which it is not clear whether the defendant was in custody, or whether he was being interrogated, or whether he waived his *Miranda* rights.⁵ In such hard cases a judge's views about *Miranda*'s wisdom will likely influence her decisions about the application of the precedent. But in the easy cases—the cases that present the *same* question that the Supreme Court decided in *Miranda*—the lower court judge is obliged by the system to decide the question as it has already been decided even if, without the constraint of precedent, she would have reached a different decision.

"Horizontal" Precedent

The constraint of precedent, at least in theory, applies horizontally as well as vertically. That is, judges are obliged to follow previous decisions of their *own* court even if, again, they disagree with those decisions. In theory, a Supreme Court Justice who in 2010 disagrees with the Court's 1973 decision in *Roe v. Wade* (1973) is obliged by what is known as the doctrine of *stare decisis*—"stand by what is decided"—to follow that decision. At least with respect to the Supreme Court, however, the data indicate that the constraint of *stare decisis* is a weak one, having little force in explaining the votes of the Justices (Brenner & Spaeth, 1995; Schauer, 2008b; Segal & Spaeth, 1996). Unlike the obligation to follow the ruling of a higher court, which is largely respected when the decision of the higher court is clear, the obligation to follow an earlier decision of the same court appears to be perceived by judges as weak.

The obligation of a judge to follow a precedent that is exactly "on point" is an important aspect of legal reasoning and the self-understanding of the legal system, but its effect is rarely seen in appellate courts. When it is clear that some dispute is the same as that which has already been decided, the dispute will usually be resolved prior to reaching the appellate court. The cases that do end up being decided on appeal, again by virtue of the selection effect, are overwhelmingly ones in which past decisions do not obviously control the current dispute but exert their influence in a less direct way. Because the idea of following precedent so pervades the legal consciousness, drawing on and arguing from past decisions even when they are not directly controlling is a ubiquitous feature of legal reasoning, argument, and decision making.

The Role of Analogical Reasoning

Using previous decisions that are not exactly like the current question in order to guide, persuade, and justify is a process that is heavily dependent on, or perhaps identical to, analogical reasoning (Spellman, 2004). Understanding the legal system's use of analogical reasoning is accordingly vitally important for understanding the methods of legal reasoning and argument. Consider, for example, the decision of the New York Court of Appeals in *Adams v. New Jersey Steamboat Company* (1896), a case frequently discussed in the literature on analogical reasoning in law (e.g., Spellman, 2010; Weinreb, 2005). The case concerned the degree of responsibility of the owner of a steamboat that contained sleeping quarters to an overnight passenger whose money had been stolen when, allegedly because of the company's negligence, a burglar broke into the passenger's stateroom. No existing legal rule controlled the case, and no previous decision had raised or decided the same question. And it turned out that two different bodies of law—two different lines of precedent—were each potentially applicable. If the law pertaining to the open sleeping compartments ("berths") in railroad cars applied, the steamboat company would not be liable to the passenger. But if the law about innkeepers' responsibility to their guests was applicable, then the passenger could recover.

The *Adams* case presents a classic case of analogical reasoning in law. Although some prominent skeptics about analogical reasoning argue that judges, like the judges in *Adams*, simply make a policy-based choice of a general rule (Greenawalt, 1992, p. 200) and mask it in the language of similarity (Alexander,

1996; Posner, 2006), such an approach is inconsistent with what we know about analogical reasoning. Applying the research on analogy to the *Adams* case, we can understand how each side was trying to get the judges to apply a different well-understood source—either the law of innkeepers or the law of railroads—to a less well-understood target—a stateroom on a steamboat.

So is a steamboat more similar to an inn or to a train? We suspect that most people would answer “train,” but that is not the relevant question. How about: Is the stateroom on a steamboat more similar to a room at an inn or to a sleeping berth on a railroad? That is a tougher question, and one might be tempted to ask (as one should when dealing with categorization generally; see Chapter 10), “Similar with respect to what?” Here the answer might be, “With respect to how much the plaintiff had the right to expect security of his possessions while he slept.” Given the situations—that one can lock one’s room at the inn and one’s stateroom on the steamboat but not one’s berth on the train; given that the room at the inn and the stateroom are more private than the sleeping berth on the train; and, perhaps, given that one paid extra for a room and a stateroom (the court’s decision did not include many details)—it is easy to argue that the steamboat and inn are similar in that the owner gives the traveler an implied guarantee that he and his possessions will be safe while sleeping.

In fact, the court applied the law of innkeepers rather than the law of railroads, and such a decision might be explained in terms of a distinction between surface and relational similarities (Holyoak & Koh, 1987). The successful analogy—between the steamboat and the inn—was not the one in which the objects were similar, but rather the one in which the legal relations between the relevant parties were similar. Developing expertise in law, which we assume the judges possessed, means seeing through the surface similarities and understanding which relational similarities matter.⁶ Note that in saying that the relevant legal category in *Adams* was a category that connects inns and steamboat accommodations (the category of those who offer sleeping accommodations, perhaps) rather than one that connects steamboats and railroads (means of transportation), the court based its categorization decision on a legal rather than a lay category.

Is this kind of reasoning substantially different between those who are legally trained and those who are not? Consider an experiment that compared law

students to undergraduates (Braman & Nelson, 2007, Exp. 2). The subjects (96 undergraduates and 77 law students) read an article summarizing the facts of a target case, but they did not know the result, and they also read one version of a potentially relevant previously decided case—which varied between subjects on two factors of possible legal relevance. The undergraduates rated the precedent as more similar to the target case than did the law students. The law students perceived similarity and difference between the cases in light of legal and not lay categories. Although the determination of similarity and difference is likely to be domain dependent (Medin, Lynch, & Solomon, 2000), it does not follow from the fact that particular similarities that are important in one domain are unimportant in another that the very process of determining similarity varies according to domain. Thus, although there may be differences between legal reasoners and ordinary reasoners, the differences, insofar as they are a function of knowledge attained in legal training and practice, may be better characterized as content based rather than process based.

Possession of legal knowledge may thus explain the difference between legally trained and non-legally trained reasoners. But given that most judges are legally trained, and given that both sides present the potentially relevant cases supporting their sides to the judges, why are there disputes over the appropriate analogy to use? The Realists would say that the judges have a desired outcome and then pick the appropriate analogies to justify their decisions. But perhaps people (and judges) choose relevant analogies (or precedents) as better or worse, applicable or inapplicable, not because of any particular desired outcome but rather because of their own preexisting knowledge and the way they frame their questions (Spellman, 2010; Spellman & Holyoak, 1992, 1996).

Fact Finding

As described earlier, an important type of decision making in legal proceedings is “fact finding” and most of the factual determinations in legal proceedings are made by judges. Many of these determinations are made in the course of preliminary proceedings and many are made in trials in which there is no jury. Yet although fact finding is done far more often by judges than by juries, most of the research about fact finding has been done on juries. One reason may be that juries feature prominently in television and movie trials, and as a result researchers may believe they are more prevalent

in nontheatrical legal proceedings than they really are (Spellman, 2006). Another reason might be that lay jurors are far more likely to resemble the typical experimental subjects used by psychology researchers. Using findings based on experiments with university undergraduates to draw conclusions about the decision-making practices of judges may involve significant problems of external validity, but the greater similarity between lay undergraduates and lay jurors significantly lessens these problems (Bornstein, 1999).

Fact Finding by Juries

Perhaps the most important dimension of jury fact finding is the way in which the information that juries receive is carefully controlled by the law of evidence. Evidence law is based on the assumption that jurors will overvalue or otherwise misuse various items of admittedly relevant information, and the rules of evidence thus exclude some relevant evidence because of a distrust of the reasoning capacities of ordinary people. There is a general rule of evidence (FRE 403) that a judge may exclude relevant evidence "if its probative value is substantially outweighed by a danger of...unfair prejudice, confusing the issues, misleading the jury..." More specifically, for example, the information that the defendant in a robbery case has committed robbery in the past is typically excluded from jury consideration, even though a rational decision maker would recognize that such evidence, even if hardly conclusive, is far from irrelevant.⁷ Similarly, the exclusion of hearsay evidence—evidence of what someone else (who is not now testifying) said was true, rather than what a witness perceived as true—is based on the notion that juries will give too much weight to what was said by a person who is not appearing in court. Yet this fear entails excluding from consideration evidence that ordinary decision makers would consider relevant to the decision to be made. In ordinary life, people rely frequently on hearsay to inform themselves about what happened and often make judgments about ambiguous or unknown current behavior based on past behavior. The fact that the law of evidence excludes so much of what figures prominently in everyday reasoning is accordingly perhaps the most important feature of evidence law.

When jurors are the fact finders, they may receive two types of instructions from the trial judge. The first are immediate instructions during the trial: to forget information they have heard or to use some information for one purpose but not for another (obvious)

other one. For example, a witness might blurt out that he knew the defendant because they had been in prison for robbery at the same time. If the defense lawyer objects and makes a motion to strike, and the judge sustains, she will immediately instruct the jury to disregard that evidence. There is much data supporting the conclusion that jurors typically do not disregard such evidence (Steblay et al., 2006). The jury is still out, however, on the question of whether it is that jurors *cannot* disregard or *choose not* to disregard. There is strong evidence that under some conditions the failure is intentional (e.g., Sommers & Kassin, 2001), but it is likely that under other conditions jurors are simply unable to disregard what they already know.

The second type of instructions comes just before jurors deliberate: They are instructed about both the content of the law specific to the case at hand and about general procedures they should use to decide the case. The latter include the mandate that they decide the case in accordance with the instructed law, and not on the basis of what they think is the right result. Thus, an important question, one about which there has been considerable research, is the extent to which jurors actually understand judge's instructions (see Diamond & Rose, 2005, and Ogloff & Rose, 2005, for reviews).

Although considerable recent efforts have aimed at making instructions more comprehensible, the research suggests that jurors typically do not understand very much of the judge's instructions, including specific instructions about elements of the crime and general instructions about the burden of proof (Ogloff & Rose, 2005). Some of this gap between instructions given and instructions comprehended may be a function of the fact that judges are more concerned with legal accuracy in language of the instruction (so the case will not be overturned on appeal) than they are with maximum comprehension by the jury. But much of the gap may follow from the difficulty that experts in general have of understanding the perspective of nonexperts in their own field.

Although jurors often do not understand the judge's instructions, that does not imply that they will deliver an erroneous verdict. It turns out that juries tend to deliver the correct verdict, at least where the measure of correctness is what the judge would have decided were there no jury. Various studies over the years, using different methodologies, have shown that a judge's and jury's decisions about the same cases are typically in accord

(see Diamond & Rose, 2005). However, each one of these studies has at least one serious methodological flaw. Still, overall, it seems that even though jurors may not appreciate the nuances of the applicable law, they are reliable in getting a general sense of who ought to prevail. (As far as we can tell, none of the research has focused explicitly on the decisions of jurors who do not understand the instructions in cases in which the justice of the situation and the law point in opposite directions, and thus it would be a mistake to assume that juror incomprehension of judicial instructions is largely inconsequential.)

That jurors who at best imperfectly understand the judge's instructions nevertheless reliably reach the correct verdict is related to what we know about how juries determine what happened. Much of the structure of a trial and much of the law of evidence is premised on an incremental and Bayesian model of fact finding, in which jurors with prior beliefs about some state of affairs adjust the probability of those beliefs upward or downward as additional pieces of evidence are presented (see Griffiths, Tenenbaum, & Kemp, Chapter 3, for a discussion of Bayesian inference and Hahn & Oaksford, Chapter 15, for applications to the jury decision making). This is a plausible model of how information is received and processed at trial, yet it is not a model that appears to track the reality of juror decision making.

The prevailing psychological model of juror decision making is the Story Model (Pennington & Hastie, 1991), which suggests that juror decision making is more holistic than incremental. The Story Model proposes that jurors evaluate the evidence based on which story (i.e., prosecution, defense, or some other) best explains all or almost all of the evidence they have heard, as opposed to making a preliminary determination on the basis of some evidence and then continually revising that determination as additional pieces of evidence are presented. In seeking the story that best explains the evidence they have heard, therefore, jurors' reasoning is largely devoted to determining which of the two (or more) competing stories at a trial is more coherent and complete. Indeed, another holistic model of reasoning, Explanatory Coherence (Thagard, 1989), has been applied to reasoning about legal cases (Simon et al, 2001; Thagard, 2003), scientific reasoning, and other types of reasoning (Thagard, 2006; Chapter 14). Thus, that these models explain ordinary reasoning as well as jury decision making provides still further support for the view that legal decision making, whether by judge or by jury, is less different

from ordinary decision making than lawyers and judges have long believed.

Note, however, that the aforementioned describes the prevailing model of *juror*, not *jury*, decision making. Often left out of studies of legal decision making is the fact that jurors, always, and judges, often (on appeal but not at trial), make their final decisions as a group. Research on group decision making is thus very relevant to legal decision making (see Salerno & Diamond, 2010).

Fact Finding by Judges

The law of evidence provides an interesting window into the legal system's traditional belief in the superior and distinctive reasoning powers of those with legal training. In countries that do not use juries there is rarely a discrete body of evidence law, and judges are comparatively free to take all relevant information into account. And in the United States, when judges sit without juries, they often tell the lawyers that many of the rules of evidence will be disregarded or interpreted loosely to allow more evidence to be considered than would be allowed were there a jury (Schauer, 2006). Underlying this practice is the belief that only those with legal training can be trusted to evaluate evidence properly (Mitchell, 2003; Robbenolt, 2005), but it turns out that underlying this belief is more unsupported faith than actual data (Spellman, 2006).

Should we expect judges to be better at fact finding than juries? There are many differences between jurors and judges (as we discuss in the next section), but there is certainly nothing about law school training that seems likely to affect this type of reasoning: It is not at all like what Professor Kingsfield had in mind with his version of training in the Socratic method whereby he would press students with question after question about the meaning and implications of the decision in a case. Perhaps, however, either judges' repeated experience listening to cases (versus jurors doing it rarely) or their desire to do the right thing in following the law (where jurors might not take that mandate as seriously) would make judges better. In terms of repeated experience, because there never is real feedback regarding what the true facts of a case were, it is doubtful that practice makes one better. And in terms of wanting to follow the law, there is research supporting the view that judges are barely better than laypeople at ignoring information they are supposed to disregard (Wistrich, Guthrie, & Rachlinski, 2005). Thus, the data that exist about judicial fact finding support the conclusion that, when acting as fact finders and not

as legal interpreters, judges are less different from lay jurors than many people—and many judges—commonly believe (Robinson & Spellman, 2005).

Judges' Expertise and the Authority of Law

There is, as we have emphasized, a running debate between the traditional and Legal Realist accounts of legal reasoning, and one way of framing the question of the distinctiveness of legal reasoning is in terms of the traditional claim that lawyers and judges are experts. That was clearly Kingsfield's claim, for example, but it leaves open the question of what kind of experts lawyers and judges might be. More particularly, is it possible that there are at least some process-based differences between legal and lay reasoning? Consider again the task of analogical reasoning in law. Perhaps lawyers and judges simply become *better* analogical reasoners by virtue of their legal training and experience. Perhaps judges, and to some extent lawyers, are experts at analogical reasoning in ways that laypeople are not.

Judges (and typically lawyers) differ from non-judges and nonlawyers on a variety of dimensions (see Stanovich, Chapter 22). On average, they have higher IQs than, say, jurors. They have more formal schooling. They may differ on some personality variables. They have chosen to go into, and stay in, the legal field. They are repeat players—doing the same thing time after time. And, as a result, they are likely motivated to “get it right,” or at least not to “get it badly wrong,” because their decisions become public and their reputations and even their jobs could be at stake. They also have their years of law school training. There is research showing that judges fall prey to the same standard reasoning biases as other mortals (e.g., anchoring, hindsight bias, etc., even when the problems are framed in a judicial context; Guthrie, Rachlinski, & Wistrich, 2001). But the research has focused on tasks other than those that are characteristically legal tasks. Maybe what Kingsfield was driving at was the notion that law students can be trained to be better at the central reasoning tasks that engage lawyers and judges.

Expertise and Analogy

That lawyers and judges are better at analogical reasoning than laypeople seems like a plausible claim, but it is not borne out by the research. Just as there are no data to support the belief that judges are expert fact finders (Robinson & Spellman, 2005) or experts at weighing evidence (Spellman, 2006), there are no data to support that judges' ability to

use analogies transcends the domains in which they normally operate. And if they are not experts at using analogies outside of the law, then the expertise they have is an expertise that comes from their legal knowledge and not from any increased ability in analogical reasoning itself. Thus, when law students in their first and third years of law school were compared to medical students and graduate students in chemistry and psychology (Lehman, Lempert, & Nisbett, 1988), the law students had initially higher scores on a verbal reasoning test (which included verbal analogies) than the others, presumably partly a function of self-selection and partly of the selection criteria of law schools. After 3 years of schooling, however, the law students showed only a statistically nonsignificant increase in verbal reasoning while the others improved to a greater extent. If these findings are generalizable, they might be thought to provide further support for the view that legal reasoning expertise, if it exists, is a content-based and not process-based expertise.

Expertise and Authority

But as described earlier, particularly in the sections on rules that give the wrong answer and on precedent, there is more to legal reasoning than using analogies. Understanding the traditional view of legal reasoning, and even the nature of law itself, requires appreciating the role that authority plays in legal decision making. Just as citizens are expected to obey the law even when they think it mistaken, so too are lawyers and judges expected to follow the legal rules and legal precedents even when they disagree with them. In this sense the law is genuinely authoritative; its force derives from its source or status rather than from its content (Hart, 1982). Just as the exasperated parent who, having failed to reason successfully with her child, asserts, “Because I said so!” law's force derives from the fact that the law says it rather than the intrinsic value of the content of what the law is saying.

The nature and power of authority has been the subject of psychological research, primarily by social psychologists, but the effect of an authority, even an impersonal authority like the law, also has cognitive dimensions. For example, authoritative sources may provide arguments and reasons that the decision maker would not otherwise have thought valid and relevant. On the other hand, sometimes an authoritative legal source will tell a decision maker to ignore what she thinks is a relevant fact (Raz, 1979), and sometimes it will tell a decision maker to consider

what she thinks is an irrelevant fact. As an example of the former: The relevant Supreme Court free speech cases make irrelevant the fact that a speaker is a member of the American Nazi Party or the Ku Klux Klan and wishes to publicly espouse Nazi or racist sentiments. The law not only demands that these factors be disregarded, but it also demands that they be disregarded even by a decision maker who disagrees with this aspect of the law. As an example of the latter: In determining whether a will is valid, a judge must determine whether the will contains the requisite signatures applied according to various other formalities, absent which the will is invalid even if there is no doubt that it represents the wishes of the deceased. And the judge is obliged to take this into account even if the judge believes it would produce an unjust outcome in this case, and even if the judge believes that the law requiring the formalities is obsolete or otherwise mistaken.

Thus, an important question is the extent to which legal decision makers can suppress their best judgment in favor of an authority with which they disagree. The traditional view of legal reasoning is that decision makers can be trained to do just that, and in fact much of the training in law school is devoted to inculcating just this kind of distinction between obedience to legal authority and taking into account that which otherwise seems morally and decisionally relevant (Schauer, 2009). Indeed, because the inherent authority of law often requires a decision maker to ignore what she thinks relevant, and consider what she believes irrelevant, it may be useful to understand part of legal reasoning as not being *reasoning* at all. It is, to be sure, decision making, but part of legal decision making is the way in which authoritative law makes legal decision makers avoid reasoning and even avoid thinking. For the legal decision maker, just like the legal subject, the authority of law is the mandate to leave the thinking and reasoning to someone else.

Are people willing and able to do that? Research by Schweitzer and colleagues (Schweitzer, Sylvester, & Saks, 2007; Schweitzer et al., 2008) indicates that law students are more willing than laypeople to follow rules even when the result produced by following a rule conflicts with the just result, suggesting that the difference between legal reasoning and ordinary reasoning may involve some process- and not content-based skills. Yet Schweitzer and colleagues also found no differences between first-year and third-year law students, possibly indicating that the process-based dimensions of legal reasoning

are more a matter of self-selection and law school admissions selection than of anything that is actually taught and learned during the study or practice of law. Perhaps, therefore, lawyers and judges are different from laypeople, but those differences may be more a function of knowledge, experience, and self-selection than of actual training in distinctively legal reasoning.

Legal Procedures

In this chapter, and throughout much of the research on legal reasoning, great emphasis has been placed on the legal decision maker. Who makes legal decisions, how might legal decision makers resemble or differ from other decision makers, and what differences, if any, might these similarities and differences make (see LeBoeuf & Shafir, Chapter 16)? But the law is not only a domain of decision makers with unique abilities, training, and experience, it is also a domain in which the procedures and structures for making decisions differ from those commonly found elsewhere. Controlling for differences in decision maker characteristics, therefore, might decision-making procedures by themselves produce important differences in the thinking and reasoning of those who are making the decisions?

The structural and procedural differences of the legal decision are manifested in numerous ways. Consider, for example, the all-or-nothing nature of much of legal decision making. Legal decisions are typically binary, with the parties winning or losing, and with legal rules or precedents being applicable or not. Probabilistic determinations are the exception and not the norm in law. A plaintiff who suffers \$100,000 damages and proves her civil case to a 60% certainty does not recover \$60,000, as expected value decision theory would suggest, but rather the entire \$100,000; and if she established the same case with 48% certainty, she would get nothing at all. A defendant who is charged with first-degree murder (which includes a finding of premeditation) cannot be found guilty of manslaughter (which does not) if he was not charged with manslaughter but the jury thinks he indeed killed the victim without the requisite intent.⁸

Similarly, it is rare for a judge to say that a rule or precedent is almost applicable or partly applicable, and even rarer for an uncertain judge, at least explicitly, to split the difference in a legal argument. There has been little research how the all-or-nothing character of legal decision making might create or explain some of the differences between legal and nonlegal decision making.

The binary character of legal decision making is merely one example of the procedural peculiarity of legal decision making, but there are many others. Judges are typically expected to provide written reasons for their decisions, but how does the requirement of formal reason-giving affect the nature of the decision?⁹ Conversely, juries are typically prohibited from explaining the reasons behind their decisions, and how might this prohibition influence their decisions? The appellate process commonly produces redundancy in decision making, but how is the decision of an appellate court influenced by the knowledge that the judge below has already reached a decision about the same questions? Finally, and perhaps most obviously, legal procedures are especially adversarial, and it would be valuable to know the extent to which decision makers—whether judges or jurors—think differently in the context of adversarial presentations than they would were the same information and arguments presented to them in a less combative or more open-ended manner. In these and other respects, it may well be that considering legal reasoning solely as a matter of content- or process-based differences (or not) is too simple, and that a psychological account of legal reasoning must be conscious of how these distinctively legal procedures and structures affect the decision makers.

Conclusion and Future Directions

We have noted at various places that most of the research on judicial decision making has been based on assumptions rather than data about the similarity between judges and lay decision makers. There are obvious problems with trying to use judges and even experienced lawyers as experimental subjects. Still, insofar as the central questions of legal reasoning from a psychological perspective are the questions of whether people can be selected (or self-select) for a certain kind of legal reasoning ability, or whether they can be trained for a certain kind of legal reasoning ability, further research on the differences between lawyers, law students, and judges, on the one hand, and laypeople, on the other, remains an essential research task.

A related agenda for research is one that would distinguish the task of fact finding from the task of interpreting, applying, and, at times, making law. The traditional claims for legal reasoning are largely about these latter functions, and thus the evaluation of the traditional claims will need to focus more on the application of rules and precedents than has thus far been the case. Only when such research

has been conducted in a systematic way will we will be able to approach an answer to the question of whether Kingsfield was right, or whether he was just the spokesman, as the more extreme of the Legal Realists claimed, for a long-standing but unsupported self-serving ideology of the legal and judicial professions.

Notes

1. In three recent Supreme Court nomination hearings, for example, now-Chief Justice Roberts insisted that Supreme Court Justices were like baseball umpires, simply calling balls and strikes with no interest in the outcome; now-Justice Sotomayor claimed that her past decisions as a judge were based solely on the law and not on her personal views, and that her future decisions would be the same; and now-Justice Kagan, even while acknowledging that Justices must exercise substantial discretion, said that good Supreme Court decisions were still based on “the law all the way down.”

2. The difference between this formulation of the Realist view and the earlier one is, did the judge first consciously decide what she wanted the outcome to be (e.g., Bush has to win in *Bush v. Gore*, 2000) and then try to justify it (strong Realism) or did the decision come unbidden, as a “hunch”? This latter version sounds a bit like the Moral Intuitionist version of moral reasoning (Haidt, 2001; see Chapter 19)—in which people make moral judgments from quick intuitions then strive to justify them—but they are different. The Moral Intuitionist view is vague about what intuitions are and how they arise; we believe that intuitions arise from knowledge, and, thus, an experienced judge’s intuition about a case will reflect her knowledge of other similar cases. She may arrive at the opinion consistent with her values not because she consciously decided which way to rule, but because her previous knowledge and beliefs gave her a justifiable intuition (Spellman 2010; see Kahneman & Klein, 2009).

3. Civil law countries are those whose legal systems emanate, for example, from the Code of Justinian in Roman times or the Napoleonic Code 2,000 years later.

4. Common-law countries include the United States, the United Kingdom, and Australia. The type of legal system tends to vary with whether the country has juries, with common-law countries using them more and civil-law countries using them less, but the covariation is not a necessary one.

5. The key is to argue that these differences make it not the “same” question.

6. Or, perhaps, because the court believed as a policy matter that they *ought* to be treated as similar and decided accordingly. Similarity judgments may be guided by pragmatic relevance (Spellman & Holyoak, 1996).

7. The rule keeping out such evidence seems concerned with people making the Fundamental Attribution Error (Ross, 1977).

8. This all-or-none nature of a probabilistic verdict provides the backdrop for pretrial settlements and plea bargains. It also affects how much money a plaintiff might ask for in a civil case and which criminal charges a prosecuting attorney will bring.

9. It is in vogue to believe that not thinking about a complex decision is best (Dijksterhuis, Bos, Nordgren & van Baaren, 2006), but there is concern about those findings (e.g., Payne, Samper, Bettman & Luce, 2008; see also McMackin & Slovic, 2000).

References

- Adams v. New Jersey Steamboat Company, 45 N.E. 369 (N.Y. 1896).
- Alexander, L. (1996). Bad beginnings. *University of Pennsylvania Law Review*, 145, 57–87.
- Bornstein, B. (1999). The ecological validity of jury simulations: Is the jury still out? *Law and Human Behavior*, 23, 75–91.
- Braman, E. (2010). Searching for constraint in legal decision making. In D. Klein & G. Mitchell (Eds.), *The psychology of judicial decision making* (pp. 203–220). New York: Oxford University Press.
- Braman, E., & Nelson, T. E. (2007). Mechanism of motivated reasoning? Analogical perception in discrimination disputes. *American Journal of Political Science*, 51, 940–956.
- Brenner, S., & Spaeth, H. J. (1995). *Stare indecisus: The alteration of precedent on the Supreme Court, 1946–1992*. New York: Cambridge University Press.
- Bush v. Gore, 531 U.S. 98 (2000).
- Coke, E. (1628). *Commentaries upon Littleton*. Birmingham, AL: Legal Classics Library (reprint of the 18th ed., Charles Butler, Ed., 1985).
- Diamond, S. S., & Rose, M. R. (2005). Real juries. *Annual Review of Law and Social Science*, 1, 255–284.
- Dijksterhuis, A., Bos, M. W., Nordgren, L. F., & van Baaren, R. B. (2006). On making the right choice: The deliberation-without-attention effect. *Science*, 311(5763), 1005–1007.
- Dworkin, R. (1986). *Law's empire*. Cambridge, MA: Harvard University Press.
- Evans, J. St. B. T. (2003). In two minds: Dual process accounts of reasoning. *Trends in Cognitive Sciences*, 7, 454–459.
- Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 59, 255–278.
- Evans, J. St. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory and Cognition*, 11, 295–306.
- Fiedler, K. (2011). Voodoo correlations – A severe methodological problem, not only in social neurosciences. *Perspectives on Psychological Science*, 6, 163–171.
- Frank, J. (1930). *Law and the modern mind*. New York: Brentano's.
- Furgeson, J. R., Babcock, L., & Shane, P. M. (2008a). Behind the mask of method: Political orientation and constitutional interpretive preferences. *Law and Human Behavior*, 32, 502–510.
- Furgeson, J. R., Babcock, L., & Shane, P. M. (2008b). Do a law's policy implications affect beliefs about its constitutionality? An experimental test. *Law and Human Behavior*, 32, 219–227.
- Galanter, M. (2004). The vanishing trial: An examination of trials and related matters in federal and state trial courts. *Journal of Empirical Legal Studies*, 1, 459–570.
- Gentner, D., & Kurtz, K. J. (2005). Relational categories. In W.-K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin* (pp. 151–175). Washington, DC: American Psychological Association.
- Greenawalt, K. (1992). *Law and objectivity*. New York: Oxford University Press.
- Guthrie, C., Rachlinski, J. J., & Wistrich, A. J. (2007). Blinking on the bench: How judges decide cases. *Cornell Law Review*, 93, 1–43.
- Guthrie, C., Rachlinski, J. J., & Wistrich, A. J. (2001). Inside the judicial mind. *Cornell Law Review*, 86, 777–830.
- Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108, 814–834.
- Hart, H. L. A. (1958). Positivism and the separation of law and morals. *Harvard Law Review*, 71, 593–629.
- Hart, H. L. A. (1982). Commands and authoritative legal reasons. In H. L. A. Hart (Ed.), *Essays on Bentham: Jurisprudence and political theory* (pp. 243–266). Oxford, England: Clarendon Press.
- Hastie, R. (1993) (Ed.). *Inside the juror: The psychology of juror decision making*. New York: Cambridge University Press.
- Holmes, O. W. (1897). The path of the law. *Harvard Law Review*, 10, 457–481.
- Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory and Cognition*, 15, 332–340.
- Holyoak, K. J., & Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General*, 128, 3–31.
- Hutcheson, J., Jr. (1929). The judgment intuitive: The function of the "hunch" in judicial decision. *Cornell Law Journal*, 14, 274–288.
- Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, 64, 515–526.
- Kennedy, D. (1986). Freedom and constraint in adjudication: A critical phenomenology. *Journal of Legal Education*, 36, 518–562.
- Kunda, Z. (1987). Motivated inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology*, 53, 636–647.
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108, 480–498.
- Lederman, L. (1999). Which cases go to trial? An empirical study of predictions of failure to settle. *Case Western Reserve University Law Review*, 49, 315–358.
- Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of graduate training on reasoning: Formal discipline and thinking about everyday-life events. *American Psychologist*, 43, 431–442.
- Llewellyn, K. (1930). *The bramble bush: On our law and its study*. New York: Columbia.
- McMackin, J., & Slovic, P. (2000). When does explicit justification impair decision making? *Applied Cognitive Psychology*, 14, 527–541.
- Medin, D. L., Lynch, E. B., & Solomon, K. O. (2000). Are there kinds of concepts? *Annual Review of Psychology*, 51, 121–147.
- Miranda v. Arizona, 384 U.S. 436 (1966).
- Mitchell, G. (2003). Mapping evidence law. *Michigan State Law Review*, 2003, 1065–1147.
- Molden, D. C., & Higgins, E. T. (2005). Motivated thinking. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 295–317). New York: Cambridge University Press.
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2, 175–220.
- Ogloff, J. R. P., & Rose, V. G. (2005). The comprehension of judicial instructions. In N. Brewer & K. D. Williams (Eds.),

- Psychology and law: An empirical perspective* (pp. 407–444). New York: Guilford Press.
- Payne, J. W., Samper, A., Bettman, J. R., & Luce, M. R. (2008). Boundary conditions on unconscious thought in complex decision making. *Psychological Science*, 19, 1118–1123.
- Pennington, N., & Hastie, R. (1991). A cognitive theory of juror decision making: The story model. *Cardozo Law Review*, 13, 519–557.
- Posner, R. A. (2006). Reasoning by analogy. *Cornell Law Review*, 91, 761–774.
- Priest, G. L., & Klein, W. (1984). The selection of disputes for litigation. *Journal of Legal Studies*, 13, 1–23.
- Raz, J. (1979). *The authority of law: Essays on law and morality*. Oxford, England: Clarendon Press.
- Redding, R. E., & Reppucci, N. D. (1999). Effects of lawyers' socio-political attitudes on their judgments of social science in legal decision making. *Law and Human Behavior*, 23, 31–54.
- Riggs v. Palmer, 22 N.E. 188 (N.Y. 1889).
- Rips, L. J. (2001). Two kinds of reasoning. *Psychological Science*, 12, 129–134.
- Robbennolt, J. (2005). Jury decisionmaking: Evaluating juries by comparison to judges: A benchmark for judging. *Florida State Law Review*, 32, 469–509.
- Robinson, P. A., & Spellman, B. A. (2005). Sentencing decisions: Matching the decisionmaker to the decision nature. *Columbia Law Review*, 105, 1124–1161.
- Roe v. Wade, 410 U.S. 113 (1973).
- Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10, pp. 173–220). New York: Academic Press.
- Salerno, J. M., & Diamond, S. S. (2010). The promise of a cognitive perspective on jury deliberation. *Psychonomic Bulletin and Review*, 17, 174–179.
- Schauer, F. (1985). Easy cases. *Southern California Law Review*, 58, 399–440.
- Schauer, F. (1991). *Playing by the rules: A philosophical examination of rule-based decision-making in law and in life*. Oxford, England: Clarendon Press.
- Schauer, F. (2006). On the supposed jury-dependence of evidence law. *University of Pennsylvania Law Review*, 155, 165–202.
- Schauer, F. (2008a). A critical guide to vehicles in the park. *New York University Law Review*, 83, 1109–1134.
- Schauer, F. (2008b). Has precedent ever really mattered in the Supreme Court? *Georgia State Law Review*, 25, 217–236.
- Schauer, F. (2008c). Why precedent in law (and elsewhere) is not totally (or even substantially) about analogy. *Perspectives on Psychological Science*, 3, 454–460.
- Schauer, F. (2009). *Thinking like a lawyer: A new introduction to legal reasoning*. Cambridge, MA: Harvard University Press.
- Schlegel, J. (1980). American legal realism and empirical social science: The singular case of Underhill Moore. *Buffalo Law Review*, 29, 195–303.
- Schweitzer, N. J., Saks, M. J., Tingen, I., Lovis-McMahon, D., Cole, B., Gildar, N., & Day, D. (2008). *The effect of legal training on judgments of rule violations*. Paper presented at the Annual Meeting of the American Psychology-Law Society, Jacksonville, FL. Retrieved September 2011, from http://www.allacademic.com/meta/p229442_index.html
- Schweitzer, N. J., Sylvester, D. J., & Saks, M. J. (2007). Rule violations and the rule of law: A factorial survey of public attitudes. *DePaul Law Review*, 47, 615–636.
- Segal, J. J., & Spaeth, H. J. (1996). The influence of stare decisis on the votes of Supreme Court Justices. *American Journal of Political Science*, 40, 971–1003.
- Segal, J. J., & Spaeth, H. J. (2004). *The Supreme Court and the attitudinal model revisited*. New York: Cambridge University Press.
- Shapiro, C. (2009). Coding complexity: Bringing law to the empirical analysis of the Supreme Court. *Hastings Law Journal*, 20, 477–537.
- Simon, D., Pham, L. B., Le, Q. A., & Holyoak, K. J. (2001). The emergence of coherence over the course of decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 1250–1260.
- Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119, 3–22.
- Sommers, S. R., & Kassin, S. M. (2001). On the many impacts of inadmissible testimony: Selective compliance, need for cognition, and the overcorrection bias. *Personality and Social Psychology Bulletin*, 27, 1368–1377.
- Spellman, B. A. (2004). Reflections of a recovering lawyer: How becoming a cognitive psychologist – and (in particular) studying analogical and causal reasoning – changed my views about the field of psychology and law. *Chicago-Kent Law Review*, 79, 1187–1214.
- Spellman, B. A. (2006). On the supposed expertise of judges in evaluating evidence. *University of Pennsylvania Law Review PENNUMbra*, 155, 1–9.
- Spellman, B. A. (2010). Judges, expertise, and analogy. In D. Klein & G. Mitchell (Eds.), *The psychology of judicial decision making* (pp. 149–163). New York: Oxford University Press.
- Spellman, B. A., & Holyoak, K. J. (1992). If Saddam is Hitler then who is George Bush? Analogical mapping between systems of social roles. *Journal of Personality and Social Psychology*, 62, 913–933.
- Spellman, B. A., & Holyoak, K. J. (1996). Pragmatics in analogical mapping. *Cognitive Psychology*, 31, 307–346.
- Spellman, B. A., Holyoak, K. J., & Morrison, R. G. (2001). Analogical priming via semantic relations. *Memory and Cognition*, 29, 383–393.
- Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.
- Steblay, N., Hosch, H. M., Culhane, S. E., & McWethy, A. (2006). The impact on juror verdicts of judicial instruction to disregard inadmissible evidence: A meta-analysis. *Law and Human Behavior*, 30, 469–492.
- Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, 12, 435–467.
- Thagard, P. (2003). Why wasn't O.J. convicted? Emotional coherence in legal inference. *Cognition and Emotion*, 17, 361–384.
- Thagard, P. (2006). Evaluating explanations in law, science, and everyday life. *Current Directions in Psychological Science*, 15, 141–145.
- Twining, W. (1973). *Karl Llewellyn and the Realist movement*. London, England: Weidenfeld & Nicolson.
- Weinreb, L. (2005). *Legal reason: The use of analogy in legal argument*. Cambridge, MA: Harvard University Press.
- Wistrich, A. J., Guthrie, C., & Rachlinski, J. J. (2005). Can judges ignore inadmissible information? The difficulties of deliberately disregarding. *University of Pennsylvania Law Review*, 153, 1251–1345.

Medical Reasoning and Thinking

Vimla L. Patel, Jose F. Arocha, and Jiajie Zhang

Abstract

The practice of medicine requires art as well as science. The latter argues for a deeper understanding of the mechanisms underlying disease processes and use of scientific evidence in making patient care decisions. The study of medical reasoning and thinking underlies much of medical cognition and has been the focus of research in cognitive science and artificial intelligence in medicine. Expertise and medical knowledge organization, the directionality of reasoning, and the nature of medical errors are intricately tied to thinking and decision-making processes in medicine. With the recent advancement of technology in medicine, technology-mediated reasoning and reasoning support systems will be a focus for future research. This chapter discusses these issues within historical and current research perspectives.

Key Words: medical reasoning, directionality of reasoning, knowledge organization, expertise, medical error, technology-mediated thinking

What Is Medical Reasoning?

Medical cognition refers to the study of cognitive structure and processes, such as perception and action, comprehension, problem solving, decision making, memory, and knowledge representation in medical practice or in tasks representative of medical practice. Medical reasoning research describes a form of qualitative and quantitative inquiry that examines the cognitive processes underlying medical decisions. Researchers study individuals at various levels of experience and in different roles in medical settings, including medical students, physicians, nurses, and biomedical scientists.

Medical problem solving, diagnostic reasoning, and decision making are all terms used in a growing body of literature that examines how clinicians understand biomedical information, solve clinical problems, and make clinical decisions. The study

of medical reasoning underlies much of medical cognition, and it has been the focus of considerable research in cognitive science and artificial intelligence as applied to medicine.

Medical reasoning involves an inferential process for making diagnostic or therapeutic decisions or understanding the pathology of a disease process. On the one hand, medical reasoning is basic to all higher level cognitive processes in medicine, such as problem solving and medical text comprehension. On the other hand, the structure of medical reasoning is itself the subject of considerable scrutiny. For example, the directionality of reasoning in medicine has been an issue of considerable controversy in medical cognition, medical education, and artificial intelligence in medicine. It is typical to partition medical reasoning into clinical and biomedical or basic science reasoning. These are some of the central themes in this chapter.

Early Research on Medical Problem Solving and Reasoning

Understanding the thought processes involved in clinical reasoning to promote more effective practices has been the subject of inquiry for more than a century (Osler, 1906). However, it was not until the development of medical cognition, a subfield of cognitive science devoted to the study of cognitive processes in medical tasks, that major scientifically based advances were made. Studies of medical cognition include analyses of performance in “real-world” clinical environments as well as in experimental settings.

There have been two primary approaches to medical reasoning: the decision-analytic approach and the information-processing approach. Decision analysis research aims to compare the performance of a physician with a mathematical model that represents the ideal of rationality and focuses on the reasoning “fallacies” and biases inherent in human clinical decision making (Leaper, Horrocks, Staniland, & De Dombal, 1972). In contrast, the information-processing approach focuses on the description of cognitive processes in reasoning tasks and the development of information-processing models of performance, typically relying on protocol analysis (Ericsson & Simon, 1993) and other techniques, such as naturalistic observation. Human information-processing research has typically focused on the individual. A dual focus on in-depth task analysis and on the study of human performance is a central feature of the information-processing approach to medical cognition, which is the focus of this section.

The information-processing view of cognitive processes came into prominence in the early 1970s, spearheaded by the immensely influential work of Newell and Simon (1972) on problem solving (see Bassok & Novick, Chapter 21). Problem solving was conceived of as search in a problem space in which a problem solver was viewed as selecting an option (e.g., a hypothesis or an inference) or performing an operation (from a set of possible operations) in moving toward a problem solution or a goal state (e.g., diagnosis or treatment plan). This conceptualization had an enormous impact in both cognitive psychology and artificial intelligence (AI). It also led to rapid advances in medical reasoning and problem-solving research, as exemplified by

the seminal work of Elstein, Shulman, and Sprafka (1978). These investigators studied the problem-solving processes of physicians, drawing on then contemporary methods and theories of cognition. Their view of problem solving had a substantial influence on both studies of medical reasoning and medical education. They were the first to use experimental methods and theories of cognitive science to investigate clinical competency. Their extensive empirical research led to the development of a model of hypothetico-deductive reasoning, which proposed that physicians reasoned by first generating and then testing a set of hypotheses to account for clinical data (i.e., reasoning from hypothesis to data). This model of problem solving had a substantial influence on studies of both medical cognition and medical education (Elstein, 2009; Elstein, Shulman, & Sprafka, 1978).

At the same time when empirical research in medical cognition was beginning, medical AI (particularly, expert systems technology) began development of methods and implementations of expert systems that mimic the way medical experts reasoned. AI in medicine and medical cognition mutually influenced each other in a number of ways, including (1) providing a basis for developing formal models of competence in problem-solving tasks; (2) elucidating the structure of medical knowledge and providing important epistemological distinctions, and (3) characterizing productive and less-productive lines of reasoning in diagnostic and therapeutic tasks. Gorry (1973) conducted a series of studies comparing a computational model of medical problem solving with the actual problem-solving behavior of physicians. This analysis provided a basis for characterizing a sequential process of medical decision making, one that differs in important respects from early diagnostic computational systems based on Bayes’ theorem (Ledley & Lusted, 1959; see Chater & Oaksford, Chapter 2; Griffiths, Tenenbaum, & Kemp, Chapter 3), which describes the relations between conditional probabilities and provides a normative rule for updating belief in light of evidence. Pauker, Gorry, Kassirer, and Schwartz (1976) capitalized on some of the insights of Gorry’s earlier work to develop the *Present Illness Program* (PIP), a program designed to take the history of a patient with edema. Several of the questions guiding this research, including the nature and organization of expert knowledge, were of central concern to both developers of medical expert systems and researchers in medical cognition. The development and refinement of the program was partially

based on studies of clinical problem solving. Medical expert consultation systems, such as Internist (Miller, Pople, & Myers, 1984) and MYCIN (Shortliffe, 1976), introduced ideas about knowledge-based reasoning strategies across a range of cognitive tasks. MYCIN, in particular, had a substantial influence on cognitive science. It contributed several advances (e.g., representing reasoning under uncertainty) in the use of production systems as a representation scheme in a complex, knowledge-based domain. MYCIN also highlighted the difference between medical problem solving and the cognitive dimensions of medical explanation. Clancey's work (Clancey & Letsinger, 1984; Clancey, 1985) in GUIDON and NEOMYCIN was particularly influential in the evolution of models of medical cognition. Clancey endeavored to reconfigure MYCIN in order to employ the system to teach medical students about meningitis and related disorders. NEOMYCIN was based on a more psychologically plausible model of medical diagnosis. This model differentiated data-directed and hypothesis-directed reasoning and separated control knowledge (rules) from the facts upon which it operated.

Feltovich, Johnson, Moller, and Swanson (1984), drawing on models of knowledge representation from medical AI, characterized fine-grained differences in knowledge organization between subjects with different levels of expertise in the domain of pediatric cardiology. These differences accounted for subjects' inferences about diagnostic cues and their evaluation of competing hypotheses. Patel and Groen (1986), incorporating distinctions introduced by Clancey, studied the knowledge-based solution strategies of expert cardiologists as evidenced by their pathophysiological explanations of a complex clinical problem. The results indicated that subjects, who accurately diagnosed the problem, employed a forward-driven reasoning strategy—using patient data to lead toward a complete diagnosis (i.e., reasoning from data to hypothesis). In contrast, subjects who misdiagnosed or partially diagnosed the patient problem used a backward-driven reasoning strategy. These research findings presented a challenge to the hypothetico-deductive model of reasoning espoused by Elstein, Shulman, and Sprafka (1978), which did not differentiate expert from nonexpert reasoning strategies.

Much of the early research in the study of reasoning in domains such as medicine was carried out in laboratory or experimental settings. More

recently, there has been a shift toward examining cognitive issues in naturalistic settings, such as medical teams in intensive care units (Patel, Kaufman, & Magder, 1996), anesthesiologists working in surgery (Gaba, 1992), nurses providing emergency telephone triage (Leprohon & Patel, 1995), and patients reasoning with the aid of technology in the health care system (Patel, Arocha, & Kushniruk, 2002). This research was informed by work in the area of dynamic decision making (Salas & Klein, 2001), complex problem solving (Frensch & Funke, 1995), human factors (Hoffman & Deffenbacher, 1992; Vicente & Rasmussen, 1990), and cognitive engineering (Rasmussen, Pejtersen, & Goodstein, 1994).

Models of Medical Reasoning

The traditional view of medical reasoning has been to treat diagnosis as similar to the classical view of scientist's task of making a discovery or engaging in scientific experimentation. This classical view of scientific reasoning makes the same assumption that diagnostic inference follows a hypothetico-deductive process of reaching conclusions by testing hypotheses based on clinical evidence. Within a cognitive perspective, as we saw previously, this view of the diagnostic process in medicine was first proposed in the influential work of Elstein, Shulman, and Sprafka (1978). However, the view of medical reasoning as hypothetico-deductive has been challenged, as we will see in this section. Similarly, the classical view of scientific reasoning has been expanded to cover multiple types of reasoning strategies.

Toward a Model of Reasoning in Medicine: Induction, Deduction, and Abduction

It is generally agreed upon that there are two basic forms of reasoning. One is deductive reasoning (see Evans, Chapter 8), which in the context of medicine consists of deriving a particular conclusion (such as a diagnosis) from a hypothesis (such as a diagnostic category or a pathophysiological process). The other form is inductive reasoning, which consists of generating a likely general conclusion (such as a diagnosis) from patient data (see Rips et al., Chapter 11). However, reasoning in the "real world" does not appear to fit neatly into any of these basic types. For this reason, a third form of reasoning has been recognized as best capturing the generation of clinical hypotheses, where deduction and induction are intermixed. This

corresponds to what Peirce (1955; see Lombrozo, Chapter 14) termed “abductive reasoning,” which in the medical context is illustrated by the clinician generating a plausible explanatory hypothesis through a process of heuristic rule utilization (Magnani, 1992, 2001).

There are different uses of the term “abductive reasoning.” In our chapter, we use it to refer to a cyclical process of generating possible explanations (i.e., identification of a set of hypotheses that are able to account for the clinical case on the basis of the available data) from a set of data and testing those explanations (i.e., evaluation of each generated hypothesis on the basis of its expected consequences) for the abnormal state of the patient at hand (Elstein, Shulman, & Sprafka, 1978; Joseph & Patel, 1990; Kassirer, 1989; Ramoni, Stefanelli, Magnani, & Barosi, 1992). Abductive reasoning is a data-driven process and also dependent on domain knowledge.¹ Within this generic framework, various models of diagnostic reasoning may be constructed. Following Patel and Ramoni (1997), we can distinguish between two major models of diagnostic reasoning: *heuristic classification* (Clancey, 1985) and *cover and differentiate* (Eshelman, 1988). However, these models can be seen as special cases of a more general model: the *select and test* model (Ramoni et al., 1992), where the processes of hypothesis generation and testing can be characterized in terms of four types of processes: abstraction, abduction, deduction, and induction. During *abstraction*, pieces of data in the data set are selected according to their relevance for the problem solution and chunked in schemas representing an abstract description of the problem at hand (e.g., abstracting that an adult male with hemoglobin concentration less than 14 g/dL is an anemic patient). Following this, hypotheses that could account for the current situation are related through a process of *abduction*, characterized by a “backward flow” of inferences across a chain of directed relations that identify those initial conditions from which the current abstract representation of the problem originates. This process provides tentative solutions to the problem at hand by way of hypotheses. For example, knowing that disease *A* will cause symptom *b*, by abduction one will try to identify the explanation for *b*, while through deduction one will forecast that a patient affected by disease *A* will manifest symptom *b*; both inferences use the same relation along two different directions. These two types of clinical reasoning in medicine are described by Patel and Ramoni (1997).

In the testing phase, hypotheses are incrementally tested according to their ability to account for the whole problem, where *deduction* serves to build up the possible world described by the consequences of each hypothesis. This kind of reasoning is customarily regarded as a common way of evaluating diagnostic hypotheses (Joseph & Patel, 1990; Kassirer, 1989; Patel, Arocha, & Kaufman, 1994; Patel, Evans, & Kaufman, 1989). As predictions are derived from hypotheses, they are matched to the case through a process of *induction*, where a prediction generated from a hypothesis can be matched with one specific aspect of the patient problem. The major feature of induction is, therefore, the ability to rule out those hypotheses whose expected consequences turn out to be not in agreement with the patient problem. This is because there is no way to logically confirm a hypothesis. We can only disconfirm or refute it in the presence of contrary evidence. This evaluation process closes the testing phase of the diagnostic cycle. Moreover, it determines which information is needed in order to discriminate among hypotheses and hence which information has to be collected.

Hypothesis Testing and Clinical Reasoning: A View From the Empirical Literature

A model such as one presented earlier can be used to explain the medical diagnostic process, while the empirical literature highlights aspects of medical reasoning that led support to the model. First, seasoned clinicians are selective in the data they collect (*abstraction*), focusing only on the data that are relevant to the generated hypotheses, while ignoring other less relevant data, as was found in the work of Patel and Groen (1986) and later supported in other studies. Successful clinicians focus on the fewest pieces of data and are better able to integrate these pieces of data into a coherent explanation for the problems (see Groves, O’Rourke, & Alexander, 2003, for an example). Second, typically physicians generate a small set of hypotheses very early in the case (*abduction*), as soon as the first pieces of data become available, as was first shown by Elstein, Shulman, and Sprafka (1978) and later corroborated by Feltovich et al.’s (1984) logical competitor set. Third, as also originally shown by Elstein, Shulman, and Sprafka (1978), physicians sometimes make use of the hypothetico-deductive process (*deduction*), which involved four stages: cue acquisition, hypothesis generation, cue interpretation, and hypothesis evaluation. Cues in the clinical case lead

to the generation of a few selected hypotheses, and then each hypothesis is evaluated for consistency with the cues (*induction*).

The empirical research has also highlighted other aspects of medical reasoning that are independent of the general model. These include the directionality of diagnostic reasoning, the role of causal and analogical reasoning in clinical problem solving, and the development of medical expertise. A series of articles (Patel, Arocha, & Kaufman, 1994, 2001; Patel, Kaufman, & Arocha, 2002) and other papers in edited volumes (Clancey & Shortliffe, 1984; Szolovits, 1982) provide summaries of the different aspects of medical cognition. To some of these issues, we turn next.

Problem-Solving Strategies: Forward-Driven and Backward-Driven Reasoning

As indicated above, the study that Patel and Groen (1986) conducted on the knowledge-based solution strategies used by expert cardiologists showing that those physicians who accurately diagnosed the problem employed a forward-driven (also called data-driven) reasoning strategy. That is, they used patient data to generate a fully correct diagnosis; whereas those who did not diagnose the problem accurately or failed to provide a correct solution to the patient problem used a backward or hypothesis-driven reasoning strategy. Such results challenged the hypothetico-deductive strategy proposed by Elstein, Shulman, and Sprafka (1978). Given the contrasting results of Patel and Groen with those of Elstein et al., it was necessary to search for an explanation that reconciled such empirical findings.

A hypothesis for reconciling these seemingly contradictory results is that forward-driven reasoning is used in clinical problems in which the physician has ample experience. In contrast, when reasoning through unfamiliar or difficult cases, physicians resort to backward-driven reasoning because their knowledge base does not support a pattern-matching process. To support this explanation, Patel, Groen, and Arocha (1990) looked for the conditions under which forward-driven reasoning breaks down. Cardiologists and endocrinologists were asked to solve diagnostic problems both in cardiology and in endocrinology. They showed that under conditions of case complexity and uncertainty, the pattern of forward-driven reasoning was disrupted. More specifically, the breakdown occurred when nonsalient cues in the case were tested for consistency against

the main hypothesis, even by subjects who had generated the correct diagnosis. The results supported previous studies in that subjects with accurate diagnoses used pure forward-driven reasoning.

If forward-driven reasoning breaks down when case complexity is introduced, then experts and novices should reason differently because what are routine cases for experts would not be so for less-than-expert subjects. Investigating clinical reasoning in a range of contexts of varying complexity (Patel, Arocha, & Kaufman, 1994; Patel & Groen, 1991a), the researchers found that novices and experts have different patterns of data-driven and hypothesis-driven reasoning. The use that experts make of data-driven reasoning, which depends on the physician possessing a highly organized knowledge base about the patient's disease (including sets of signs and symptoms). Furthermore, due to their extensive knowledge base and the high-level inferences they make, experts typically skip steps in their reasoning. In contrast, because of their lack of substantive knowledge or their inability to distinguish relevant from irrelevant knowledge, less-than-expert subjects (novices and intermediates) used more hypothesis-driven reasoning, resulting often in very complex reasoning patterns. Similar patterns of reasoning have been found in other domains (Larkin et al., 1980).

The fact that experts and novices reason differently suggests that they might reach different conclusions (e.g., decisions or understandings) when solving medical problems. Although data-driven reasoning is highly efficient, it is often error prone in the absence of adequate domain knowledge, since there are no built-in checks on the legitimacy of the inferences that a person makes. Pure data-driven reasoning is only successful in constrained situations, where one's knowledge of a problem can result in a complete chain of inferences from the initial problem statement to the problem solution. In contrast, hypothesis-driven reasoning is slower and requires high memory load, because one has to keep track of goals and hypotheses. It is therefore most likely to be used when domain knowledge is inadequate or the problem is complex. Hypothesis-driven reasoning is an example of a *weak method* of problem solving in the sense that it is used in the absence of relevant prior knowledge and when there is uncertainty about problem solutions (see Bassok & Novick, Chapter 21). In problem-solving terms, strong methods engage knowledge, whereas weak methods refer to general, knowledge-independent

strategies. “Weak” does not necessarily imply ineffectual in this context.

Studies also showed that data-driven reasoning could break down due to uncertainty (Patel, Groen, & Arocha, 1990). These conditions include the presence of “loose ends” in explanations, where some particular piece of information remains unaccounted for and isolated from the overall explanation. Loose ends trigger explanatory processes that work by hypothesizing a disease, for instance, and trying to fit the loose ends within it, in a hypothesis-driven reasoning fashion. The presence of loose ends may foster learning, as the “reasoner” searches for an explanation for them. For example, a medical student or a physician may encounter a sign or a symptom in a patient problem and look for information that may account for the finding, by searching for similar cases seen in the past, reading a specialized medical book, or consulting a domain expert.

However, in some circumstances, the use of data-driven reasoning may lead to a heavy cognitive load (see van Merriënboer & Sweller, 2005, for a recent review). For instance, when students are given problems to solve while they are being trained in the use of problem-solving strategies, the situation produces a heavy load on cognitive resources, which may diminish students’ ability to focus on the task. The reason is that students have to share cognitive resources (e.g., attention, memory) between learning the problem-solving method and learning the content of the material. Research by Sweller (1988) suggested that when subjects use a strategy based on data-driven reasoning, they are more able to acquire a schema for the problem. In addition, other characteristics associated with expert performance were observed, such as a reduced number of moves to the solution. However, when subjects used a hypothesis-driven reasoning strategy, their problem-solving performance suffered.

The Role of Similarity of Cases in Diagnostic Reasoning

The fact that physicians make use of forward-driven reasoning in routine cases suggests a type of processing that is fast enough to be able to lead to the recognition of a set of signs and symptoms in a patient and generate a diagnosis based on such recognition. Most often this has been interpreted as a type of specific-to-general reasoning (e.g., reasoning from an individual case to a clinical schema or prototype). However, consistent with the model of abductive reasoning, some philosophers (Schaffner,

1986) and empirical researchers (Norman & Brooks, 1997) have supported an alternative hypothesis, which postulates specific-to-specific reasoning. That is, experts may use knowledge of specific instances (e.g., particular patients with specific disease presentations) to interpret particular cases, rather than relying only on general clinical knowledge (Kassirer & Kopelman, 1990). Similarity-based clinical reasoning has been informed by models of categorization, where the process of interpreting information, such as a clinical case, can be made by matching the case to a general, abstract category (e.g., a schema). However, in similarity-based reasoning a diagnosis may be reached through a process of matching the current case to similar cases seen in the past (Norman, Young, & Brooks, 2007).

Brooks and colleagues (Brooks, Norman, & Allen, 1991; Norman & Brooks, 1997) investigated clinicians’ use of specific instances in order to compare and interpret a current clinical case. In these studies, mostly involving visual diagnosis based on X-rays, dermatological slides, and electrocardiograms, they showed that specific similarity to previous cases accounts for about 30% of diagnoses made. Furthermore, errors made by experts in identifying abnormalities in images were affected by the prior history of the patient. That is, if the prior history of the patient contained a possible abnormality, expert physicians more often identified those abnormalities in the images even when none were there, which was interpreted as the effect of specific past cases on the interpretation of the current case.

Norman and colleagues (Norman & Brooks, 1997; Norman, Young, & Brooks, 2007) have argued against the hypothesis that expert physicians always diagnose clinical cases by “analyzing” signs and symptoms and developing correspondences between those signs, symptoms, and diagnoses, in the manner that traditional cognitive research in medical reasoning suggests. They propose instead that medical diagnosis is often “nonanalytic.” By this, they mean that diagnostic reasoning is driven by the overall similarity between a previous case and the current case: A case previously seen in medical practice is retrieved unanalyzed from memory and compared to the current case (a kind of exemplar-based or case-based reasoning).

This discussion has its counterpart in the psychology of categorization, where two accounts have been proposed: Either categorization works by relying on prototypes or on exemplars (see Rips et al., Chapter 11). Exemplar-based thinking is certainly

an important process in human cognition. There is ample evidence of conditions where reasoning by analogy to previous cases is used (Gentner & Holyoak, 1997; Holyoak & Thagard, 1997; see Holyoak, Chapter 13). Furthermore, given the complexity of natural reasoning in a highly dense knowledge domain such as medicine, it is very likely that more than one type of reasoning is actually employed. Seen in this light, the search for a single manner in which clinicians diagnose clinical problems may not be a reasonable goal. The inherent adaptability of humans to different kinds of knowledge domains, situations, problems, and cases may call for the use of a variety of reasoning modes, which is what, after all, the notion of abductive medical reasoning has tried to describe (Patel & Ramoni, 1997; see Lombrozo, Chapter 14). Similar conclusions appear to have been developed in the dual-process theory of reasoning (Evans, 2008; see Evans, Chapter 8). Thus, it seems that alongside a rule-based, more effortful prototype reasoning, clinical reasoning also allows for exemplar-based reasoning. However, it should be noted that not all reasoning from instances/analogs is nonanalytic. For example, analogical inferences may involve causal analysis of source analog in which similarity between particulars may be the main cognitive mechanism.

Reasoning and the Nature of Medical Knowledge

A reason for the variety of modes of reasoning used in actual diagnostic problems may be found in the inherent organization of medical knowledge. The prevalent view in the philosophy of medicine (Blois, 1988) has been that medical knowledge has an extremely complex organization, requiring the use of different reasoning strategies than those used in other formal scientific disciplines, such as physics. According to Blois, disciplines such as physics, chemistry, and some subfields of biology, are *horizontally* organized, where these domains are characterized by the construction of causal relations among concepts and by the application of general principles to specific instances (Blois, 1988). Furthermore, Blois asserts that such scientific fields are organized in a hypothetico-deductive manner such that particular cases are explained from first principles, where understanding the causal mechanisms plays a major role. In contrast, in medicine, reasoning from first principles does not play such an important role. He argued that reasoning in medicine instead requires

vertical thinking. In this view, in the medical disciplines, notably clinical medicine, reasoning by analogy from case to case plays a more important role than reasoning causally from an understanding of the mechanism of disease. Based on this distinction, Blois argued that reasoning in the physical sciences and reasoning in the biomedical sciences are of different kinds.

Schaffner (1986) has argued that theories in the physical sciences can be conceptualized as a “deductive systematization of a broad class of generalizations under a small number of axioms” (Schaffner, 1986, p. 69), but that such characterization cannot be applied to the biomedical sciences. The latter are characterized by what Schaffner (1986, p. 68) calls “a series of overlapping interlevel temporal models.” Different level of aggregation (e.g., biochemical, cellular, organ), each with its own temporal development (e.g., some deterministic, others probabilistic, others still random) makes it difficult to develop deductive structures to explain biomedical reality. Rather, theories are based on the familiarization with shared exemplars to a much greater degree than is the case in the physical sciences. In biomedical research, an organism such as a *Drosophila*, for instance, is used as an exemplar embodying a given disease mechanism, which by analogy applies to other organisms, including humans. In the clinical sciences, the patient is seen as an exemplar to which generalizations based on multiple overlapping models are applied from disease mechanisms (e.g., physiological, biochemical, pathological) and from the population of similar patients (e.g., typical diagnostic categories described in clinical medicine).

A reason for the peculiar nature of biomedical knowledge may be that the field consists of two very different categories of knowledge: clinical knowledge, including knowledge of disease processes and associated findings; and basic science knowledge, incorporating subject matters such as biochemistry, anatomy, and physiology. Basic science or biomedical knowledge is supposed to provide a scientific foundation for clinical reasoning. The conventional view is that basic science knowledge can be seamlessly integrated into clinical knowledge, analogously to the way that learning the rules of the road can contribute to one's mastery of driving a car. A particular piece of biomedical knowledge could be automatically elicited in a range of clinical contexts and tasks in more or less the same fashion. However, the integration of these two types of knowledge has been difficult to demonstrate in empirical research.

Knowledge Organization and Changes in Directionality

Following Blois (1988) and Schaffner (1986), it can be argued that the way medical knowledge is organized can be a determinant factor explaining why experts do not use the hypothetico-deductive method of reasoning. Maybe the medical domain is too messy to allow neat partitioning and deductive use of reasoning strategies. Although the theory of reasoning in medicine is basically a theory of expert knowledge, reaching the level of efficient reasoning of the expert clinician reflects the extended continuum of training and levels of reasoning performance (Chi, Bassok, Lewis, Reiman, & Glaser, 1989; Dufresne, Gerace, Hardiman, & Mestre, 1992). This continuum is related to the nature of medical knowledge and its acquisition. Changes have been described in this process that serve to characterize the various phases medical trainees go through to become expert clinicians.

An important characteristic of this process is the *intermediate effect*. Although it generally seems reasonable to assume that performance will improve with training or time on task, there appear to be particular transitions in which subjects exhibit a certain drop in performance. This is an example of what is referred to in the developmental literature as non-monotonicity (Strauss & Stavy, 1982), and it has also been observed in skill acquisition (Robertson & Glines, 1985). The result is a learning curve or developmental pattern that is shaped like either a U or an inverted U, as illustrated in Figure 37.1. In the development of medical expertise, the performance

of intermediates reflects the degradation in reasoning that results from the acquisition of knowledge during a period when such knowledge is not well organized and irrelevant associations abound in the learner's knowledge base. In contrast, the novice's knowledge base is sparse, containing very few associations, whereas the expert's knowledge base has been pruned of the irrelevancies that characterize intermediates.

It should be noted that not all intermediate performance is nonmonotonic; for example, on some global criteria such as diagnostic accuracy there appears to be a steady improvement (Patel & Groen, 1991b). However, this increase in accuracy seems to be linked to a change in diagnostic strategy from hypothesis-driven reasoning to data-driven reasoning (Coderre, Mandin, Harasym, & Fick, 2003), and better organized knowledge in memory. Thus, knowledge acquisition results in changes in knowledge organization, which result in changes in the strategies used to solve clinical problems, which in turn lead to greater diagnostic accuracy. What the intermediate effect tells us is that the progression from a novice to an expert does not necessarily follow a linear improvement in performance, but a U-shaped pattern, where performance drops before it gets better. This is illustrated in Figure 37.1, where medical students' recall of a clinical problem (after reading it once) showed an inverted U-shaped pattern with the third-year students recalling case information the most (Patel & Groen, 1991b). Their inability to separate relevant from irrelevant information was consistently validated under various task conditions. This

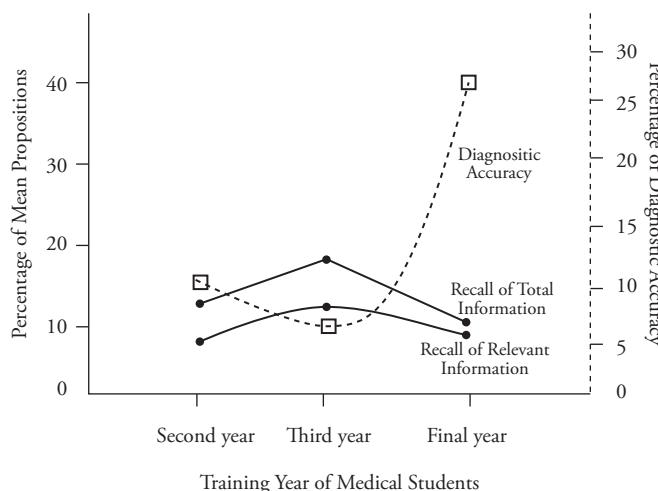


Fig. 37.1 The “intermediate effect.” Recall of total and relevant information as a function of expertise, showing inverted U-shaped pattern, with intermediates recalling more relevant and irrelevant information. The performance level as a function of diagnostic accuracy and expertise shows a U-shaped phenomenon, where performance drops at the intermediate level (third year) before it gets better with training.

inability reflects in a decrease in diagnostic accuracy. Novices do not know enough, while experts know enough and have a well-developed schema for the problem to make accurate diagnoses. Intermediate-level subjects remember the most, but their knowledge is not organized in such a fashion that allows them to generate accurate diagnoses.

Now let us look at some details of the intermediate effect. The intermediate effect occurs with many tasks and at various levels of expertise. The tasks vary from recall of clinical cases (Claessen & Boshuizen, 1985), explanation of clinical problems (Arocha & Patel, 1995), to generating laboratory data (Patel, Groen, & Patel, 1997). For instance, when asked to recall clinical data from case descriptions, intermediates tend to recall more irrelevant information (Schmidt & Boshuizen, 1993) or generate more irrelevant hypotheses to account for a clinical case than novices or experts (Arocha & Patel, 1995). The phenomenon may be due to the fact that intermediates have acquired an extensive body of knowledge but have not yet reorganized this knowledge in a functional manner. Thus, intermediate knowledge has a sort of network structure that results in considerable search, which makes it more difficult for intermediates to set up structures for rapid encoding and selective retrieval of information (Patel & Groen, 1991a). In contrast, expert knowledge is finely tuned to perform various tasks and experts can readily filter out irrelevant information using their hierarchically organized schemata. The difference is reflected both in the structural organization of knowledge and the extent to which it is proceduralized to perform different tasks.

Schmidt and Boshuizen (1993) reported that intermediate nonmonotonicity recall effects disappear by using short exposure times (about 30 seconds), which suggests that under time-restricted conditions, intermediates cannot engage in extraneous search. While a novice's knowledge base is likely to be sparse and an expert's knowledge base is intricately interconnected, the knowledge base of an intermediate possesses many of the pieces of knowledge in place but lacks the extensive connectedness of an expert. Until this knowledge becomes further consolidated, the intermediate is more likely to engage in unnecessary search. Whether this knowledge, painfully acquired during medical training, is really necessary for clinical reasoning has been a focus of intensive research and great debate. If expert clinicians do not explicitly use underlying biomedical knowledge, does that mean that

it is not necessary? Or could it be simply the case that this knowledge remains "dormant" until it is really needed? This raises an important question of whether expert medical knowledge when applied to solving clinical problems is "deep" (e.g., making use of causal pathophysiological or basic science knowledge in clinical problem solving) or "shallow" (e.g., relying on the use of associations between disease categories and clinical manifestations of those categories, without considering the mechanism of the disease).

Causal Reasoning in Medicine

The differential role of basic science knowledge (e.g., physiology and biochemistry) in solving problems of varying complexity and the differences between subjects at different levels of expertise (Patel, Arocha, & Kaufman, 1994) have been a source of controversy in the study of medical cognition (Patel & Kaufman, 1994), as well as in medical education and artificial intelligence. As expertise develops, the disease knowledge of a clinician becomes more dependent on clinical experience, and clinical problem solving is increasingly guided by the use of exemplars and analogy and becomes less dependent on a functional understanding of the system in question. However, an in-depth conceptual understanding of basic science plays a central role in reasoning about complex problems and is also important in generating explanations and justifications for decisions.

Researchers in artificial intelligence were confronted with similar problems in extending the utility of systems beyond their immediate knowledge base. Biomedical knowledge can serve different functional roles depending on the goals of the system. Most models of diagnostic reasoning in medicine can be characterized as being "shallow." For instance, a "shallow" medical expert system (e.g., MYCIN and INTERNIST) reasons by relating observations to intermediate hypotheses that partition the problem space, and further associating intermediate hypotheses with diagnostic hypotheses. This is consistent with the way physicians appear to reason. There are, however, other medical reasoning system models that propose a "deep" mode of reasoning as a main mechanism. Chandrasekeran, Smith, and Sticklen (1989) characterize a deep system as one that embodies a causal mental model of bodily function and malfunction, similar to the models used in qualitative physics (Bobrow, 1985). Systems such as MDX-2 (Chandrasakeran

et al., 1989) or QSIM (Kuipers, 1987) have explicit representations of structural components and their relations, the functions of these components (in essence their purpose), and their relationship to behavioral states.

To become licensed physicians, medical trainees undergo a lengthy training process that includes learning of biomedical sciences, including biochemistry, physiology, and anatomy. There is an apparent contradiction between this type of training and the seeming absence of “deep” biomedical knowledge being used during expert medical reasoning. To account for such apparent inconsistency, Boshuizen and Schmidt (1992) proposed a learning mechanism, *knowledge encapsulation*. Knowledge encapsulation is a learning process that involves the organization of biomedical propositions and their interrelations in associative clusters, under a small number of higher level clinical propositions with the same explanatory power. Through exposure to clinical training, biomedical knowledge presumably becomes integrated with clinical knowledge. Biomedical knowledge can be “unpacked” when needed, but it is not used as a first line of explanation.

Boshuizen and Schmidt (1992) cite a wide range of clinical reasoning and recall studies that support this kind of learning process. Of particular importance is the well-documented finding that with increasing levels of expertise, physicians produce explanations at higher levels of generality, using fewer and fewer biomedical concepts while producing consistently accurate responses. The intermediate effect can also be accounted for as a stage in the encapsulation process in which a trainee’s network of knowledge has not yet become sufficiently differentiated, thus resulting in more extensive processing of information.

Knowledge encapsulation provides an appealing account of a range of developmental phenomena in the course of acquiring medical expertise. However, the integration of basic science in clinical knowledge is a rather complex process, and encapsulation is likely to be only part of the knowledge development process. Basic science knowledge is likely to play a different role in different clinical domains. For example, clinical expertise in perceptual domains, such as dermatology and radiology, necessitates a relatively robust model of anatomical structures that is the primary source of knowledge for diagnostic classification. In other domains, such as cardiology and endocrinology, basic science knowledge has a

more distant relationship with clinical knowledge. The misconceptions evident in physicians’ biomedical explanations would argue against their having well-developed encapsulated knowledge structures in which basic science knowledge could easily be retrieved and applied when necessary.

The results of research into medical problem solving are consistent with the idea that clinical medicine and biomedical sciences constitute two distinct and not completely compatible “worlds,” with distinct modes of reasoning and quite different ways of structuring knowledge (see Patel et al., 1994). Clinical knowledge is based on a complex taxonomy that relates disease symptoms to underlying pathology. In contrast, biomedical sciences are based on general principles defining chains of causal mechanisms. Thus, learning to explain how a set of symptoms is consistent with a diagnosis may be very different from learning how to explain what causes a disease. Although basic science knowledge and clinical knowledge can be seen as worlds apart, this does not mean that basic biomedical knowledge is completely erased from an expert’s disease schemata; rather, as suggested by the hypothesis of encapsulated knowledge, the two types of knowledge are not integrated into a whole (Rikers, Schmidt, & Moulaert, 2005). Thus, when physicians are trying to solve clinical cases at the limits of their expertise, they may resort to using biomedical knowledge to make sense of the clinical information and tie different concepts together, but in normal practice they seem to primarily use either schemata or exemplars that map directly onto clinical cases.

This “two worlds” position, although not incompatible with the notion of encapsulation, is supported in the study of conceptual understanding in biomedicine. The progression of mental models (White & Frederiksen, 1990) has been used as an alternative framework for characterizing the development of conceptual understanding in biomedical contexts. Mental models are dynamic knowledge structures that are composed to make sense of experience and to reason across spatial and/or temporal dimensions. An individual’s mental models provide predictive and explanatory capabilities of the function of a given system. White and Frederiksen employed the progression of mental models to explain the process of understanding increasingly sophisticated electrical circuits. This notion can be used to account for differences between novices and experts in understanding circulatory physiology, describing misconceptions, and explaining the

generation of spontaneous analogies in causal reasoning (Kaufman, Patel, & Magder, 1996).

Running a mental model is a potentially powerful form of reasoning, but it is also cognitively demanding. It may require an extended chain of reasoning and the use of complex representations. It is apparent that skilled individuals learn to circumvent long chains of reasoning and chunk or compile knowledge across intermediate states of inference (Chandrasekaran, 1994; Newell, 1990). This results in shorter, more direct, inferences that are stored in long-term memory and are directly available to be retrieved in the appropriate contexts. Chandrasekaran (1994) refers to this sort of knowledge as *compiled causal knowledge*. This term refers to knowledge of causal expectations that people compile directly from experience and partly by chunking results from previous problem-solving endeavors. The goals of the individual and the demands of recurring situations largely determine which pieces of knowledge get stored and used. When a physician is confronted with a similar situation, she can employ this compiled knowledge in an efficient and effective manner. The development of compiled knowledge is an integral part of the acquisition of expertise.

The idea of compiling declarative knowledge bears a certain resemblance to the idea of knowledge encapsulation. However, the claim differs in two important senses. First, the process of compiling knowledge is not one of subsumption or abstraction, and the original knowledge (uncompiled mental model) may no longer be available in a similar form (Kuijpers & Kassirer, 1984). Second, rather than being prestored unitary structures, mental models are composed dynamically out of constituent pieces of knowledge. The use of mental models is somewhat opportunistic and the learning process is less predictable. The compilation process can work in reverse as well. That is to say, discrete cause-and-effect relationships can be integrated into a mental model as a student reasons about complex physiological processes.

Errors in Medical Reasoning and Decision Making

One critical step toward understanding the cognitive mechanisms of performance problems in medical reasoning is to categorize medical errors along cognitively meaningful dimensions. Reason (1990) defines human error as a failure of achieving the intended outcome in a planned sequence of mental or physical activities. He divides human errors into

two major categories: (1) slips that result from the incorrect execution of a correct action sequence, and (2) mistakes that result from the correct execution of an incorrect action sequence. Norman's theory of action (Norman, 1986) decomposes a human activity into seven stages. Based on Reason's definition of human error and Norman's action theory, Zhang, Patel, Johnson, and Shortliffe (2004) developed a cognitive taxonomy. In this cognitive taxonomy, goal and intention mistakes are mistakes about declarative knowledge (knowledge about factual statements and propositions), such as "Motrin is a pain reliever and fever reducer." Action specification mistakes and action execution mistakes are mistakes about procedural knowledge (knowledge about procedures and rules), such as "give 1 tsp Motrin to a child per dosage up to 4 times a day if the child has fever or toothache and the weight of the child is 24–35 lb."

Goal mistakes and intention mistakes are caused by many complex factors such as incorrect knowledge, incomplete knowledge, misuse of knowledge, biases, faulty heuristics, and information overload. For example, neglect of base rate information could result in incorrect diagnosis of a disease. This is a well-documented finding in human decision making (Tversky & Kahneman, 1974). As another example, the goal of "treating the disease as pneumonia" could be a mistake if it is a misdiagnosis based on incomplete knowledge (e.g., without X-ray images). Intention mistakes can be caused by similar factors. For example, a physician treating a patient with oxygen set the flow control knob between 1 and 2 liters per minute, not realizing that the scale numbers represented discrete rather than continuous settings. As a result, the patient did not receive any oxygen. This is a mistake due to incomplete knowledge.

The use of heuristics is another common source of goal and intention mistakes. A heuristic that is often used is the reliance on disease schemata during clinical diagnosis. Disease schemata are knowledge structures that have been formed from previous experience with diagnosing diseases and contain information about relevant and irrelevant signs and symptoms. When physicians and medical students diagnose patients, they tend to rely on their schemata and base their reasoning on the apparent similarity of patient information with these schemata, instead of a more objective analysis of patient data. The schemata that are used in diagnosis often guide future reasoning about the patient, affecting what

tests are run and how data are interpreted. Arocha and Patel (1995) found that medical students and trainees maintained their initial hypotheses, even if subsequent data were contradictory. Therefore, if the initial hypothesis is wrong, errors in diagnosis and treatment are likely to occur. Preliminary presentation of the patient (e.g., signs and symptoms), then, becomes very important, because it can strongly suggest hypotheses that will be strongly held (i.e., lead to the use of schemata).

Medical Reasoning and Decision Research

Decision making is central to medical activity. Although health care professionals are generally highly proficient decision makers, their erroneous decisions have become the source of considerable public scrutiny, as described in three National Academy Press reports (Institute of Medicine, 1999, 2001, 2004).

Decisions involve the application of reasoning to select some course of action that achieves the desired goal. Hastie (2001) has identified three components of decision making: (a) choice options and courses of actions; (b) beliefs about objective states, processes, and events in the world, including outcomes states and means to achieve them; and (c) desires, values, or utilities that describe the consequences associated with the outcomes of each action-event combination. In this process, reasoning plays a major role. Research on hypothesis testing in the medical domain has shown the pervasiveness of confirmation bias (Patel, Groen, & Norman, 1993), which is evidenced by the generation of a hypothesis and the subsequent search for evidence consistent with the hypothesis, often leading to the failure to adequately consider alternative diagnostic possibilities. This bias may result in a less-than-thorough investigation with possible adverse consequences for the patient. A desire to confirm one's preferred hypothesis may moreover contribute to increased inefficiency and costs by leading to orders for additional laboratory tests that will do little to revise one's opinion, providing largely redundant data (Chapman & Elstein, 2000).

In natural setting of medicine, team decision making is the rule rather than the exception. Naturalistic decision making (NDM) is concerned with the study of cognition in "real-world" work environments that are often dynamic (e.g., rapidly changing; see Klein, Orasanu, Calderwood, & Zsambok, 1993). The majority of this research combines conventional protocol analytic methods with innovative

methods designed to investigate reasoning and behavior in realistic settings (Rasmussen, Pejtersen, & Goodstein, 1994; Woods, 1993). The study of decision making in the work context necessitates an extended cognitive science framework beyond typical characterizations of knowledge structures, processes, and skills, including modulating variables such as stress, time pressure, and fatigue as well as communication patterns that affect team performance.

Among the issues investigated in NDM are understanding how decisions are jointly negotiated and updated by participants differing substantially in their areas of expertise (e.g., pharmacology, respiratory medicine); how the complex communication process in these settings occurs; what role technology plays in mediating decisions and how it affects reasoning; and what the sources of error are in the decision-making process.

Research by Patel, Kaufman, and Magder (1996) studied decision making in a medical intensive care unit (ICU) with the objective of describing jointly negotiated decisions, communication processes, and the development of expertise. Intensive care decision making is characterized by a rapid, serial evaluation of options leading to immediate action, where reasoning is schema driven in a forward direction toward action with minimal inference or justification. However, when patients do not respond in a manner consistent with the original hypothesis, then the original decision comes under scrutiny. This strategy can result in a brainstorming session in which the team retrospectively evaluates and reconsiders the decision and considers possible alternatives. Various patterns of reasoning are used to evaluate alternatives in these brainstorming sessions, including probabilistic reasoning, diagnostic reasoning, and biomedical causal reasoning. Supporting decision making in clinical settings necessitates an understanding of communication patterns.

Another type of decision making that has attracted attention recently in medical cognition research is opportunistic decision making during task transitions. In medical settings (critical care in particular), the environment is stressful, time sensitive, interruption laden, and information rich. Franklin and colleagues (Franklin et al., 2011) studied how decisions are made when clinicians finished one task and transitioned to another task. The authors showed that the clinicians made two kinds of decisions: *planned*, in which they actively selected their next activity, and *opportunistic*, in which their next action in a series of behaviors was not determined

by protocol but rather arose through unanticipated conditions (i.e., time, proximity or resources).

In summary, although traditional approaches to decision making have viewed decisions as choosing among known alternatives, real-world decision making involves reasoning that is constrained by dynamic factors, such as stress, time pressure, risk, and team interactions. Examining medical reasoning in social and collaborative settings is even more important when information technologies are part of the ebb and flow of clinical work.

Reasoning and Medical Education

The failures and successes of reasoning strategies and skills can be traced back to their sources: education. There is evidence suggesting that the way physicians reason results from the way they have been educated. Medical education in North America as well as in the rest of the world has followed a similar path: from practice-based training to an increasingly scientific training.

Motivated by the increasing importance of basic scientific knowledge in the context of clinical practice, problem-based learning (PBL) was developed on the premise that not only should physicians possess the ordered and systematic knowledge of science, but they should also *think* like scientists during their practices. Consistent with this idea, an attempt was made to teach hypothetico-deductive reasoning to medical students, as a way to provide an adequate structure to medical *problem solving*. After all, this was the way scientists were supposed to make discoveries.

However, based on cognitive research in other knowledge domains, some researchers argued that the hypothetico-deductive method might not be the most efficient way of solving clinical problems. To investigate how the kind of training medical students receive affected their reasoning patterns, Patel, Groen, and Norman (1993) looked at the problem-solving processes of students in two medical schools with different modes of instruction: classical and problem-based curricula. They found that students in the problem-based curriculum reasoned in a way that was consistent with their training methods, showing a preponderance of hypothetico-deductive reasoning and extensive elaborations of biomedical information. The PBL students used hypothesis-driven reasoning—from the hypothesis to explain the patient data—while non-PBL students used mainly data-driven reasoning—from data toward the hypothesis. In explaining clinical

cases, PBL students produced extensive elaborations using detailed biomedical information, which was relatively absent from non-PBL students' explanations. It appears that PBL promotes the activation and elaboration of prior knowledge. However, these elaborations resulted in the generation of errors.

Patel, Arocha, and Lecissi (2001) also investigated the effects of non-PBL curricula on the use and integration of basic science and clinical knowledge and its impact on diagnostic explanation. The results showed that biomedical and clinical knowledge are not integrated and that very little biomedical information is used in routine problem-solving situations. There is significant use of expert-like data-driven strategies, however, in non-PBL students' explanations. The use of biomedical information increases when the clinical problems are complex; at the same time, hypothesis-driven strategies replace the data-driven strategies. Similar results were found in other professional domains such as law (Krieger, 2004).

Students from a PBL school integrated the two types of knowledge (biomedical sciences and clinical), and in contrast to the non-PBL students they spontaneously used biomedical information in solving even routine problems. It appeared that for students in the non-PBL curriculum, the clinical components of the problems are treated separately from the biomedical science components. The two components of the problem analysis seem to be viewed as serving different functions. However, when needed, the biomedical knowledge is utilized and seems to act as "glue" that ties the two kinds of information together.

In the PBL curriculum, the integration of basic science and clinical knowledge is so tight that students appear unable to separate the two types of knowledge. As a result, PBL students generate unnecessarily elaborate explanations, leading to errors of reasoning. PBL seems to promote a type of learning in which basic biomedical knowledge becomes so tightly tied to specific clinical problem types that it becomes difficult to decouple this knowledge in context in order to transfer to a new situation (Anderson, Reder, & Simon, 1996; Holyoak, 1985).

This outcome is consistent with how biomedical information is taught in the classroom in PBL schools, namely, by encouraging use of the hypothetico-deductive method, resulting in a predominantly backward-directed mode of reasoning. Elaborations are accompanied by a tendency to generate errors of scientific fact and flawed patterns of explanation,

such as circular reasoning. Even though a student's explanation may be riddled with bugs and misconceptions, their harmful effects may be dependent on the direction of reasoning. If they reason forward, then they are likely to view their existing knowledge as adequate. In this case, misconceptions may be long-lasting and difficult to eradicate. If they reason backward, misconceptions might best be viewed as transient hypotheses which, in the light of experience, are either refuted or else modified to form the kernel of a more adequate explanation. Interestingly, differences in the patterns of reasoning acquired in both PBL and non-PBL medical curricula are found to be quite stable, even after the students have completed medical school and are in residency training programs (Patel, Arocha, & Lecissi, 2001; Patel & Kaufman, 2001).

Instruction that emphasizes decontextualized abstracted models of phenomena has not yielded much success in medicine or in other spheres of science education. It is widely believed that the amount of transfer will be a function of the overlap between the original domain of learning and the target domain (Holyoak, 1985). The PBL's emphasis on real-world problems represents a very good source of transfer to clinical situations. However, it is very challenging to create a problem set that most effectively embodies certain biomedical concepts while maximizing transfer. Knowledge that is overly contextualized can actually reduce transfer.

Technology-Mediated Reasoning and Thinking in Medicine

All technologies mediate human performance. Technologies, whether they be computer-based or in some other form, transform the ways individuals and groups behave. They do not merely augment, enhance, or expedite performance, although a given technology may do all of these things. The difference is not one of quantitative change, but one that is qualitative in nature. Technology, tools, and artifacts not only enhance people's ability to perform tasks but also change the way they perform tasks. In cognitive science, this ubiquitous phenomenon is called the representational effect, which refers to the phenomenon that different representations of a common abstract structure can generate dramatically different representational efficiencies, task complexities, and behavioral outcomes (Zhang & Norman, 1994; see Markman, Chapter 4).

One approach to the study of how technology mediates thinking and reasoning is to consider

technology as external representations (Zhang, 1997; Zhang & Norman, 1994). External representations are the knowledge and structure in the environment, as physical symbols, objects, or dimensions (e.g., written symbols, beads of abacuses, dimensions of a graph), and as external rules, constraints, or relations embedded in physical configurations (e.g., spatial relations of written digits, visual and spatial layouts of diagrams, physical constraints in abacuses). The information in external representations can be picked up, analyzed, and processed by perceptual systems alone, although the top-down participation of conceptual knowledge from internal representations can sometimes facilitate or inhibit the perceptual processes.

The mediating role of technology can be evaluated at several levels of analysis. For example, electronic medical records alter the practice of individual clinicians in significant ways, as discussed later. Changes to an information system substantially impacts organizational and institutional practices from research to billing to quality assurance. Even the introduction of patient-centered medical records early in the 20th century necessitated changes in hospital architecture and considerably effected work practices in clinical settings. Salomon, Perkins, and Globerson (1991) introduced a useful distinction in considering the mediating role of technology on individual performance, the *effects with* technology and the *effects of* technology. The former is concerned with the changes in performance displayed by users while equipped with the technology. For example, when using an effective medical information system, physicians should be able to gather information more systematically and efficiently. In this capacity, medical information technologies may alleviate some of the cognitive load associated with a given task and permit them to focus on higher order thinking skills, such as hypothesis generation and evaluation. The *effects of* technology refer to enduring changes in general cognitive capacities (knowledge and skills) as a consequence of interaction with a technology. For example, frequent use of information technologies may result in lasting changes in medical decision-making practices even in the absence of the system.

In several studies involving the mediating role of technology in clinical practice, Patel and her colleagues (Patel et al., 2000) observed the change of thinking and reasoning patterns caused by a change in methods of writing patient records from paper records to electronic medical records (EMRs). They

found that before using EMRs, physicians focused on exploration and discovery, used complex propositions, and tended to use data-driven reasoning. After using EMRs, which have structured data, physicians focused on problem solving, used simple propositions, and tended to use problem-directed and hypothesis-driven reasoning. The change of behavior caused by the use of EMRs remained when the physicians went back to paper records, showing the enduring effects of technology on human reasoning in medicine.

As the basis for many medical decisions, diagnostic reasoning requires the collection, understanding, and use of many types of patient information, such as history, lab results, symptoms, prescriptions, images, and so on. It is affected by not just the expertise of the clinicians but also by the way the information is acquired, stored, processed, and presented. If we consider clinicians as rational decision makers, the format of a display, as long as it contains the same information, should not affect the outcome of the reasoning and decision-making process. But the formats of displays do affect many aspects of clinicians' task performance. Recently there have been several studies of how different displays of information in EMRs affect clinicians' behavior. Three major types of displays have been studied: source based, time based, and concept based. Source-based displays organize medical data by the sources of the data, such as encounter notes, lab reports, medications, lab results, radiology imaging and report, and physical exams. Time-based displays organize medical data as a temporal history of patient data. Concept-based displays organize medical data by clinically meaningful concepts or problems. In this case all data related to a specific problem are displayed together. For example, if a patient has symptoms such as coughing, chest pain, and fever, the lab results, imaging reports, prescriptions, assessments, and plans are displayed together. Zeng et al. (2002) found that different displays were useful for different tasks. For example, source-based displays aid clinicians in retrieving information for a specific test or procedure from a specific department, whereas concept-based displays aid the search of information related to a specific disease.

Medication management is always a challenging task for patients. How drug information is displayed can affect patients' understanding of the drug information, the way they manage their medication, and ultimately their health outcomes. For example, Day (1988) demonstrated how alternative

representations should be used for different purposes for improving medication scheduling. With the rapid growth of computer-based information systems, people are interacting more and more with computer-generated health information displays. These displays need to be designed to effectively and accurately generate the information that people need for informed reasoning.

Conclusions and Future Directions

Advances in cognitive science have made significant contributions to investigations into the process of medical reasoning. However, there are a number of issues that could benefit from research, especially involving multidisciplinary approaches. At the theoretical level, an important issue for future theoretical and empirical development involves the integration of the diverse modes of medical reasoning. This includes the integration of the expertise and the decision-making approaches to expert cognition (see Kahneman & Klein, 2009), which has direct implications for establishing in more detail the role that intuitive and conscious reasoning play in the different medical tasks and in the commission of errors. Some attempts have been recently made (Croskerry, 2009; Norman, 2009; Norman & Eva, 2010) to clarify their role, particularly through application of the dual-process view of thinking and reasoning (Evans, 2008) to clinical problem solving. As some have pointed out (Glöckner & Witteman, 2010), both types of processes may involve different mechanisms, which remain to be empirically and theoretically established. In this regard, the medical environment may be used as a test bed for investigating these fundamental problems in an applied real-world arena.

As such an applied area, reasoning in a medical context involves complex compositions of various populations, settings, and a high degree of uncertainty (as in critical care environments). Compounded with constraints imposed by resource availability, these factors lead to increased use of heuristic strategies. The utility of heuristics lies in limiting the extent of purposeful search through data sets. By reducing redundancy, such heuristics have substantial practical value, as a significant part of a physician's cognitive effort is based on heuristic thinking. However, the use of heuristics introduces considerable bias in medical reasoning, often resulting in a number of conceptual and procedural errors. These include misconceptions about laws governing probability, instantiation of general

rules to a specific patient at the point of care, incorrect use of prior probabilities and actions, as well as false confirmation of hypotheses.

With the increasing move of cognitive studies toward investigations in real-world settings, complex phenomena can be better studied. The constraints of laboratory-based work make it impossible to fully capture the dynamics of real-world problems. This issue is particularly salient in rapidly changing critical care environments. Studies of thinking and reasoning in medicine, which focus on medical errors and technology-mediated cognition, are increasingly paying attention to dimensions of medical work in clinical settings. The recent concern with understanding and reducing medical errors provides an opportunity for cognitive scientists to use cognitive theories and methodologies to address the pressing, practical problem of managing these errors under conditions of urgency and multitasking. A trend in health care, spurred partly by the advent of information technologies that foster communication, is the development of health care systems that are increasingly multidisciplinary and collaborative, often spanning geographic regions. Increasing costs of health care and rapid knowledge growth have accelerated the trend towards collaboration of health care professionals to share knowledge and skills. Comprehensive patient care necessitates the communication of health care providers in different medical domains, thereby optimizing the use of their expertise. Research on reasoning will need to continue to move toward a distributed model of cognition. This model will include a focus on both socially shared and technology-mediated reasoning. Finally, today medical applications from the domain of artificial intelligence span from molecular medicine to organizational aspects of work domain, the role of modeling human reasoning and cognitive science will need to be reevaluated. Modeling and reasoning will play a significant role as we strive to build successful systems to support reasoning processes (Patel, Shortliffe, Stefanelli, et al., 2009).

Acknowledgment

This chapter is dedicated to the memory of a close friend and a colleague, late Marco F. Ramoni, who was passionate about understanding the nature of reasoning in biomedicine, as he was passionate about life in general.

Note

1. In contrast, early, precognitive, accounts of medical diagnosis (Rimoldi, 1961) described it in a way that was independent of the underlying structure of the domain knowl-

edge, through the application of general reasoning strategies. These accounts simply made the assumption that some domain of knowledge existed and that all the hypotheses needed to explain a problem were available when the diagnostic process starts. For instance, skill in diagnosing patients was assumed to be the result of the application of knowledge-independent strategies to clinical problems (e.g., by following the scientific method).

References

- Anderson, J. R., Reder, L. M., & Simon, H. A. (1996). Situated learning and education. *Educational Researcher*, 25(4), 5–11.
- Arocha, J. F., & Patel, V. L. (1995b). Novice diagnostic reasoning in medicine: Accounting for evidence. *Journal of the Learning Sciences*, 4(4), 355–384.
- Blois, M. S. (1988). Medicine and the nature of vertical reasoning. *New England Journal of Medicine*, 318, 847–851.
- Bobrow, D. G. (Ed.). (1985). *Qualitative reasoning about physical systems*. Cambridge, MA: MIT Press.
- Boshuizen, H. P. A., & Schmidt, H. G. (1992). On the role of biomedical knowledge in clinical reasoning by experts, intermediates, and novices. *Cognitive Science*, 16(2), 153–184.
- Brooks, L. R., Norman, G. R., & Allen, S. W. (1991). Role of specific similarity in a medical diagnostic task. *Journal of Experimental psychology: General*, 120(3), 278–287.
- Chandrasekaran, B. (1994). The functional representation and causal process. In M. Yovitz (Ed.), *Advances in computing* (pp. 73–143). New York: Academic Press.
- Chandrasekaran, B., Smith, J. W., & Sticklen, J. (1989). Deep models and their relation to diagnosis. *Artificial Intelligence in Medicine*, 1, 29–40.
- Chapman, G. B., & Elstein, A. S. (2000). Cognitive processes and biases in medical decision making. In G. B. Chapman & F.A. Sonnenberg (Eds.), *Decision making in health care: Theory, psychology, and applications* (pp. 183–210). New York: Cambridge University Press.
- Chi, M. T. H., Bassok, M., Lewis, M. W., Reiman, P., & Glaser, R. (1989). Self explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13, 145–182.
- Claessen, H. F. A., & Boshuizen, H. P. A. (1985). Recall of medical information by students and doctors. *Medical Education*, 19, 61–67.
- Clancey, W. J. (1985). Heuristic classification. *Artificial Intelligence*, 27, 289–350.
- Clancey, W. J., & Letsinger, R. (1984). NEOMYCIN: Reconfiguring a rule-based expert system for application to teaching. In W. J. Clancey & E. H. Shortliffe (Eds.), *Readings in medical artificial intelligence: The first decade* (pp. 361–381). Reading, MA: Addison-Wesley.
- Clancey, W. J., & Shortliffe, E. H. (1984) (Eds.). *Readings in medical artificial intelligence: The first decade*. Reading, MA: Addison-Wesley.
- Coderre, S., Mandin, H., Harasym, P. H., & Fick, G. H. (2003). Diagnostic reasoning strategies and diagnostic success. *Medical Education*, 37(8), 695–703.
- Croskerry, P. (2009). Clinical cognition and diagnostic error: Applications of a dual process model of reasoning. *Advances in Health Sciences Education*, 14, 27–35.
- Day, R. S. (1988) Alternative representations. *The Psychology of Learning and Motivation*, 22, 261–305.

- Dowie, J., & Elstein, A. S. (Eds.). (1988). *Professional judgment: A reader in clinical decision making*. Cambridge, England: Cambridge University Press.
- Dufresne, R. J., Gerace, W. J., Hardiman, P. T., & Mestre, J. P. T. (1992). Constraining novices to perform expertlike problem analyses: Effects on schema acquisition. *The Journal of the Learning Sciences*, 2(3), 307–331.
- Elstein, A. S. (2009). Thinking about diagnostic thinking: A 30-year perspective. *Advances in Health Science Education, Theory and Practice*, 14(Suppl. 1), 7–18.
- Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning*. Cambridge, MA: Harvard University Press.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.
- Eshelman, L. (1988). MOLE: A knowledge acquisition tool for Cover-and-Differentiate systems. In S. C. Marcus (Ed.), *Automating knowledge acquisition for expert systems* (pp. 37–80). Boston, MA: Kluwer.
- Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgement, and social cognition. *Annual Review of Psychology*, 59, 255–278.
- Felтовich, P. J., Johnson, P. E., Moller, J. H., & Swanson, D. B. (1984). LCS: The role and development of medical knowledge in diagnostic expertise. In W. J. Clancey & E. H. Shortliffe (Eds.), *Readings in medical artificial intelligence: The first decade* (pp. 275–319). Reading, MA: Addison-Wesley.
- Frensch, P. A., & Funke, J. (1995). *Complex problem solving: The European perspective*. Hillsdale, NJ: Erlbaum.
- Franklin, A., Liua, Y., Li, Z., Nguyen, V., Johnson, R.R., Robinson, D., Okafor, N., King, B., Patel, V.L., Zhang, J. (2011). Opportunistic decision making and complexity in emergency care, *Journal of Biomedical Informatics*, 44(3), 469–476.
- Gaba, D. M. (1992). Dynamic decision-making in anesthesiology: Cognitive models and training approaches. In D. A. Evans & V. L. Patel (Eds.), *Advanced models of cognition for medical training and practice* (pp. 123–147). New York: Springer-Verlag.
- Gentner, D., & Holyoak, K. J. (1997). Reasoning and learning by analogy: Introduction. *American Psychologist*, 52(1), 32–34.
- Glöckner, A., & Witteman, C. (2010). Beyond dual-process models: A categorisation of processes underlying intuitive judgement and decision making. *Thinking and Reasoning*, 16, 1–25.
- Gorry, G. A. (1973). Computer-assisted clinical decision-making. *Methods of Information in Medicine*, 12(1), 45–51.
- Groves, M., O'Rourke, P., & Alexander, H. (2003). The clinical reasoning characteristics of diagnostic experts. *Medical Teacher*, 25(3), 308–313.
- Hastie, R. (2001). Problems for judgment and decision making. *Annual Review of Psychology*, 52, 653–683.
- Hoffman, R., & Deffenbacher, K. (1992). A brief history of applied cognitive psychology. *Applied Cognitive Psychology*, 6(1), 1–48.
- Holyoak, K. J. (1985). The pragmatics of analogical transfer. *The Psychology of Learning and Motivation*, 19, 59–87.
- Holyoak, K. J., & Thagard, P. (1997). The analogical mind. *American Psychologist*, 52(1), 35–44.
- Institute of Medicine. (1999). *To err is human: Building a safer health system*. Washington, DC: National Academy Press.
- Institute of Medicine. (2001). *Crossing the quality chasm: A new health system for the 21st century*. Washington, DC: National Academy Press.
- Institute of Medicine. (2004). *Patient safety*. Washington, DC: National Academy Press.
- Joseph, G. M., & Patel, V. L. (1990). Domain knowledge and hypothesis generation in diagnostic reasoning. *Medical Decision Making*, 10(1), 31–46.
- Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, 64, 515–526.
- Kassirer, J. P. (1989). Diagnostic reasoning. *Annals of Internal Medicine*, 110(11), 893–900.
- Kassirer, J. P., & Kopelman, R. I. (1990). Diagnosis and the structure of memory. 2. Exemplars, scripts, and simulation. *Hospital Practice (Office Edition)*, 25(11), 29–33, 36.
- Kaufman, D. R., Patel, V. L., & Magder, S. (1996). The explanatory role of spontaneously generated analogies in reasoning about physiological concepts. *International Journal of Science Education*, 18, 369–386.
- Klein, G. A., Orasanu, J., Calderwood, R., & Zsambok, C. E. (Eds.). (1993). *Decision making in action: Models and methods*. Norwood, NJ: Ablex.
- Krieger, S. H. (2004). Domain knowledge and the teaching of creative legal problem solving, *Clinical Law Review*, 11, 149–207.
- Kuipers, B. (1987). Qualitative simulation as causal explanation. *IEEE Transactions on Systems, Man, and Cybernetics*, 17, 432–444.
- Kuipers, B., & Kassirer, J. P. (1984). Causal reasoning in medicine: Analysis of a protocol. *Cognitive Science*, 8(4), 363–385.
- Larkin, J. H., McDermott, J., Simon, H. A., & Simon, D. P. (1980). Expert and novice performance in solving physics problems. *Science*, 208, 1335–1342.
- Leaper, D. J., Horrocks, J. C., Staniland, J. R., & De Dombal, F. T. (1972). Computer-assisted diagnosis of abdominal pain using “estimates” provided by clinicians. *British Medical Journal*, 4(836), 350–354.
- Ledley, R. S., & Lusted, L. B. (1959). Reasoning foundations of medical diagnosis. *Science*, 130, 9–21.
- Leprohon, J., & Patel, V. L. (1995). Decision-making strategies for telephone triage in emergency medical services. *Medical Decision Making*, 15(3), 240–253.
- Magnani, L. (1992). Abductive reasoning: Philosophical and educational perspectives in medicine. In D. A. Evans & V. L. Patel (Eds.), *Advanced models of cognition for medical training and practice* (pp. 21–41). Berlin, Germany: Springer-Verlag.
- Magnani, L. (2001). *Abduction, reason, and science: Processes of discovery and explanation*. London: Kluwer Academic-Plenum.
- Miller, R. A., Pople, H. E., & Myers, J. D. (1984). Internist-I, an experimental computer-based diagnostic for general internal medicine. In W. J. Clancey & E. H. Shortliffe (Eds.), *Readings in medical artificial intelligence: The first decade* (pp. xvi, p. 512). Reading, MA: Addison-Wesley.
- Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
- Norman, D. A. (1986). Cognitive engineering. In D. A. Norman & S. W. Draper (Eds.), *User centered system design: New perspectives on human-computer interaction* (pp. 31–61). Hillsdale, NJ: Erlbaum.

- Norman, G. R. (2009). Dual processing and diagnostic errors. *Advances in Health Sciences Education*, 14, 37–49.
- Norman, G. R., & Brooks, L. R. (1997). The non-analytical basis of clinical reasoning. *Advances in Health Sciences Education*, 2(2), 173–184.
- Norman, G. R., & Eva, K. W. (2010). Diagnostic error and clinical reasoning. *Medical Education*, 44, 94–100.
- Norman, G. R., Brooks, L. R., & Allen, S. W. (1989). Recall by expert medical practitioners and novices as a record of processing attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(6), 1166–1174.
- Norman, G., Young, M., & Brooks, L. (2007). Non-analytical models of clinical reasoning: The role of experience. *Medical Education*, 41(12), 1140–1145.
- Osler, W. (1906). *Aequanimitas. With other addresses to medical students, nurses and practitioners of medicine*. Philadelphia, PA: Blakiston's.
- Patel, V. L., Arocha, J. F., & Kaufman, D. R. (1994). Diagnostic reasoning and expertise. *Psychology of Learning and Motivation*, 31, 137–252.
- Patel, V. L., Arocha, J. F., & Kaufman, D. R. (2001). A primer on aspects of cognition for medical informatics. *Journal of the American Medical Informatics Association*, 8(4), 324–343.
- Patel, V. L., Arocha, J. F., & Kushniruk, A. W. (2002). Patients' and physicians' understanding of health and biomedical concepts: relationship to the design of EMR systems. *Journal of Biomedical Informatics*, 35(1), 8–16.
- Patel, V. L., Arocha, J., & Lecissi, M. (2001). Impact of undergraduate medical training on housestaff problem solving performance: Implications for health education in problem-based curricula. *Journal of Dental Education*, 65(11), 1199–1218.
- Patel, V. L., Evans, D. A., & Kaufman, D. R. (1989). Cognitive framework for doctor-patient communication. In D. A. Evans & V. L. Patel (Eds.), *Cognitive science in medicine: Biomedical modeling* (pp. 257–312). Cambridge, MA: MIT Press.
- Patel, V. L., & Groen, G. J. (1986). Knowledge-based solution strategies in medical reasoning. *Cognitive Science*, 10, 91–116.
- Patel, V. L., & Groen, G. J. (1991a). The general and specific nature of medical expertise: A critical look. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 93–125). New York: Cambridge University Press.
- Patel, V. L., & Groen, G. J. (1991b). Developmental accounts of the transition from medical student to doctor: Some problems and suggestions. *Medical Education*, 25(6), 527–535.
- Patel, V. L., Groen, G. J., & Arocha, J. F. (1990). Medical expertise as a function of task difficulty. *Memory and Cognition*, 18(4), 394–406.
- Patel, V. L., Groen, G. J., & Norman, G. R. (1993). Reasoning and instruction in medical curricula. *Cognition and Instruction*, 10(4), 335–378.
- Patel, V. L., Groen, C., & Patel, Y. (1997). Cognitive aspects of clinical performance during patient workup: The role of medical expertise. *Advances in Health Sciences Education*, 2(2), 95–114.
- Patel, V. L., & Kaufman, D. R. (1994). On poultry expertise, precocious kids, and diagnostic reasoning. *Academic Medicine*, 69(12), 971–972.
- Patel, V. L., & Kaufman, D. R. (2001, Feb 02). Medical education isn't just about solving problems. *The Chronicle of Higher Education*, p. B12.
- Patel, V. L., Kaufman, D. R., & Arocha, J. F. (2002). Emerging paradigms of cognition in medical decision-making. *Journal of Biomedical Informatics*, 35, 52–75.
- Patel, V. L., Kaufman, D. R., & Magder, S. (1996). The acquisition of medical expertise in complex dynamic decision-making environments. In K. A. Ericsson (Ed.), *The road to excellence: The acquisition of expert performance in the arts and sciences, sports and games* (pp. 127–165). Hillsdale, NJ: Erlbaum.
- Patel, V. L., & Ramoni, M. F. (1997). Cognitive models of directional inference in expert medical reasoning. In P. J. Feltovich & K. M. Ford (Eds.), *Expertise in context: Human and machine* (pp. 67–99). Cambridge, MA: MIT Press.
- Patel, V. L., Shortliffe, E. H., Stefanelli, M., Szolovits, P., Berthold, M. R., Bellazzi, R., & Abu-Hanna, A. (2009). The coming of age of artificial intelligence in medicine. *Artificial Intelligence in Medicine*, 46, 5–17.
- Pauker, S. G., Gorry, G. A., Kassirer, J. P., & Schwartz, W. B. (1976). Towards the simulation of clinical cognition. *American Journal of Medicine*, 60, 981–996.
- Peirce, C. S. (1955). Abduction and induction. In C. S. Peirce & J. Buchler (Eds.), *Philosophical writings of Peirce* (pp. 150–156). New York: Dover.
- Ramoni, M. F., Stefanelli, M., Magnani, L., & Barosi, G. (1992). An epistemological framework for medical knowledge based system. *IEEE Transactions on Systems, Man, and Cybernetics*, 22, 1361–1375.
- Rasmussen, J., Pejtersen, A. M., & Goodstein, L. P. (1994). *Cognitive systems engineering*. New York: Wiley.
- Reason, J. T. (1990). *Human error*. Cambridge, England: Cambridge University Press.
- Rikers, R. M. J. P., Schmidt, H. G., & Moulaert, V. (2005). Biomedical knowledge: Encapsulated or two worlds apart? *Applied Cognitive Psychology*, 19(2), 223–231.
- Rimoldi, H. J. A. (1961). The test of diagnostic skills. *Journal of Medical Education*, 36, 73–79.
- Robertson, R. J. & Glines, L. A. (1985). The phantom plateau returns. *Perceptual and Motor Skills*, 61, 55–64.
- Salas, E., & Klein, G. A. (2001). *Linking expertise and naturalistic decision making*. Mahwah, NJ: Erlbaum.
- Salomon, G., Perkins, D., & Globerson, T. (1991). Partners in cognition: Extending human intelligence with intelligent technologies. *Educational Researcher*, 20(4), 2–9.
- Schaffner, K. F. (1986). Exemplar reasoning about biological models and diseases: A relation between the philosophy of medicine and philosophy of science. *Journal of Medicine and Philosophy*, 11, 63–80.
- Schmidt, H. G., & Boshuizen, H. P. A. (1993). On the origin of intermediate effects in clinical case recall. *Memory and Cognition*, 21, 338–351.
- Shortliffe, E. H. (1976). *Computer-based medical consultations, MYCIN*. New York: Elsevier.
- Strauss, S., & Stavy, R. (1982). *U-shaped behavioral growth*. New York: Academic Press.
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12, 257–285.
- Szolovits, P. (Ed.). (1982). *Artificial intelligence in medicine* (Vol. 51). Boulder, CO: Westview Press.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131.
- van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, 17(2), 147–176.

- Vicente, K. J., & Rasmussen, J. (1990). The ecology of human-machine systems. II: Mediating "direct perception" in complex work domains. *Ecological Psychology*, 2, 207–250.
- White, B. Y., & Frederiksen, J. R. (1990). Causal model progressions as a foundation for intelligent learning environments. In W. J. Clancey & E. Soloway (Eds.), *Artificial Intelligence: Artificial intelligence and learning environments* [Special issue]. 42(1), 99–157.
- Woods, D. D. (1993). Process-tracing methods for the study of cognition outside of the experimental psychology laboratory. In G. A. Klein, J. Orasanu R. Calderwood, & C. E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 228–251). Norwood, NJ: Ablex.
- Zeng, Q., Cimino, J. J., & Zou, K. H. (2002). Providing concept-oriented views for clinical data using a knowledge-based system: an evaluation. *Journal of the American Medical Informatics Association*, 9(3), 294–305.
- Zhang, J. (1996). A representational analysis of relational information displays. *International Journal of Human-Computer Studies*, 45(1), 59–74.
- Zhang, J. (1997). The nature of external representations in problem solving. *Cognitive Science*, 21(2), 179–217.
- Zhang, J., & Norman, D. A. (1994). Representations in distributed cognitive tasks. *Cognitive Science*, 18, 87–122.
- Zhang, J., Patel, V. L, Johnson, T. R., & Shortliffe, E. H. (2004). A cognitive taxonomy of medical errors, *Journal of Biomedical Informatics*, 37, 193–204.

Thinking in Business

Jeffrey Loewenstein

Abstract

Thinking and reasoning enter into the practice of business in limitless ways. Basic and applied cognition researchers stand to gain by increasing their awareness of each other's work. To foster interaction, this chapter provides an introduction to applied cognitive research in the main areas of business academia, which collectively map out a large range of business practice: consumer behavior within marketing, organizational behavior within management, management science within operations, behavioral finance within finance, and behavioral accounting within accounting.

Key Words: cognition, business, consumer behavior, organizational behavior, management science, behavioral finance, behavioral accounting

The practice of business is enormously variable. The marketer influencing a customer's purchase, the executive negotiating a deal, the manager coordinating the production of goods, the analyst reviewing company performance, and the accountant trying to make the numbers add up are all engaging in aspects of the practice of business. To address the variety, I will provide introductions to research on thinking occurring in the context of marketing, management, operations, finance, and accounting activity. These are arguably the major functional areas of business, and not coincidentally, the major areas of business academia. Accordingly, this chapter organizes research on thinking in business by the domain of business practice, with introductions to each corresponding area of business academia.

Business schools foster a substantial amount and range of applied research on thinking and reasoning. This work, like all applied research, makes two main contributions back to the basic research areas on which it draws: It translates basic research to advance real-world practice, and it provides contexts that raise new questions and phenomena for

basic research. This review aims to foster exchange between applied research on thinking and reasoning in business academia and basic research on thinking and reasoning.

There is important work on all main areas of business practice, but there is more work in some areas than others. I discuss the areas in rough descending order of the amount of current research on thinking in the different academic areas: marketing, management, operations, finance, and accounting. Each academic area operates relatively independently, with its own journals and professional associations, roughly like the fields of psychology, anthropology, economics, sociology, and linguistics operate within social science—there is no single journal or forum in which all areas converge, just numerous overlapping subfields. Each area of business academia has a segment engaged in behavioral research, within which scholars are engaged in research on thinking: consumer behavior (within marketing), organizational behavior (within management), management science (despite the name, it lies within operations), behavioral finance, and behavioral accounting.

Of all cognitive topics studied within business academia, decision-making research is by far the most common. The book based on Herbert Simon's dissertation, *Administrative Behavior* (1947), was particularly influential in making decision making a central behavioral topic in business academia. Because of Simon's (e.g., 1955) persuasive arguments for bounded rationality, meaning that human cognitive processing is limited, Simon also characterized decision making as a behavioral concern. To avoid confusion, I note that it was Simon's work from the 1940s and 1950s that was most influential in business academia; his work from the 1960s and 1970s that had such a large influence on cognitive science, such as *Sciences of the Artificial* (1969) and, with Allan Newell, *Human Problem Solving* (1972), is less widely known in business academia, just as *Administrative Behavior* is less widely known in cognitive science. As behavioral decision-making advanced and gained legitimacy in both psychology and economics due to the work of Kahneman and Tversky and others (Kahneman, Slovic, & Tversky, 1982), and as business schools expanded their hiring of faculty from psychology departments, decision making became an established topic across business school departments. It also has more ties between basic and applied researchers than any other topic in thinking and reasoning research (see LeBoeuf & Shafir, Chapter 16; Camerer & Smith, Chapter 18).

Accordingly, in this review I devote less space to decision-making research than it warrants based on its current volume and more space to other thinking and reasoning research, such as work on learning, categorization, expertise, creativity, and group cognition. These topics are central concerns of thinking and reasoning, and they have particular importance to core business concerns such as innovation—the development and implementation of new products and processes. Broadening the thinking and reasoning topics under discussion is an attempt to help foster new ties and to give a richer view of thinking in business. In addition, I emphasize research published in applied journals published by business academic societies and organizations, rather than the work on thinking and reasoning published in social psychology journals and industrial/organizational psychology journals. That work is no less important or insightful, but because the business outlets and papers are less likely to be familiar to basic researchers studying thinking and reasoning, there is greater potential for novel cross-fertilization.

Marketing: Consumer Behavior

Businesses need to sell their products and services, so a major concern of business is shaping how consumers make purchases, use products, and think about brands. Consumer behavior researchers study these questions and generate more psychological research on individual thinking and reasoning than researchers in any other area of business academia. As a simple indication of the role of cognition research in consumer behavior, the *Handbook of Consumer Psychology* (Haugtvedt, Kerr, & Kardes, 2008) dedicates about half of its 1,200 pages to reviewing information processing and social cognition research (see also, e.g., Loken, 2006). Most consumer behavior research on thinking and reasoning is experimental. There is also mathematical and computational modeling, field survey research, observations of consumer activity, examinations of archival measures of consumer activity, and some qualitative research.

Decision Making

Consumer purchasing is a decision-making activity. For example, one prominent feature of the consumer decision-making context (as any walk through a grocery store or time spent shopping online will make apparent) is a concern for how people make decisions when confronted by large numbers of options (e.g., Broniarczyk, Hoyer, & McAlister, 1998; Hoch, Bradlow, & Wansink, 1999). This work has identified tensions between larger numbers of options providing an opportunity to maximize fit to consumer preferences and making the choice process more difficult and more likely to spur regret (Broniarczyk, 2008; Chernev, forthcoming).

An intriguing development in this area is to consider the relationship among items in the choice set. Drawing on the contrast between alignable and nonalignable differences (Markman & Gentner, 1993; see Holyoak, Chapter 13), it is possible for items in a choice set to differ in ways that are alignable along a single dimension (e.g., increasing power or size) or in ways that are nonalignable (each member possessing a distinct categorical feature). People are more likely to make a choice from a large choice set if its members are all alignable (Gourville & Soman, 2005). Thus, the alignability of the options in a choice set is a moderator of choice set size on people's likelihood of making rather than avoiding making a decision. This work is consistent with research on other means of making large assortments easier to navigate, such as having detailed preferences (Chernev, 2003) due to expertise.

The context of consumer behavior also makes salient, unlike most decision-making research, that the process of deciding can itself offer pleasure and value. Individual preferences for, say, seeking out variety (McAlister & Pessemier, 1982) can guide the construction of choice sets even for purchases that are for others (Chowdhury, Ratneshwar, & Desai, 2009). Thus, one's own pleasure or annoyance at the decision-making process, in addition to considerations of the outcome, guides decision making.

More generally, decision making is an activity, and the many goals decision makers have as they engage in that activity guide the choices that result. For example, reviewing an array of research, Bettman, Luce, and Payne (2008) argue that four broad goals recur across a wide array of consumer decisions, if not decisions more broadly: concerns for accuracy (or decision quality), minimizing effort, minimizing negative emotion, and increasing the justification for the decision. Any of these broad goals can dominate decision making, and people can trade off value across goals.

People also have goals related to specific products that influence how they evaluate options. For example, in one series of experiments, people were asked to imagine wanting to buy a fast computer, and then read about a series of computers with features that strongly or weakly supported that product-related goal or were irrelevant to the goal (Meyvis & Janiszewski, 2002). It appears that people note whether features support the goal they have in mind when evaluating the product's features, and because irrelevant features do not support the goal, they temper judgments (a dilution effect). For instance, participants' explicitly noting whether the features were relevant or irrelevant did not eliminate the dilution effect, but having participants read the features before learning the goal did eliminate the effect. Thus, goals guide how people frame or encode information, which in turn guides their evaluations.

One real-world implication of the presence of multiple goals is that people sometimes exhibit inaccurate judgments of key features of choice options, such as price. For example, one study found that about half of the customers stopped and questioned in grocery stores could not name the price of the item they had just put in their shopping carts (Dickson & Sawyer, 1990). Strikingly, their price estimates deviated from the true prices by an amount that was, on average, as large as the range of prices in the product category.

Learning and Expertise

Another important concern is how consumers learn about products, services, and brands. Accordingly, there is a long-standing interest in consumer expertise (Alba & Hutchinson, 1987). This means that in addition to a concern for explaining a particular choice, another reason to examine consumers' goals is to examine the effect of goals on what consumers learn about products. For example, one study examined exemplar learning by low and high domain knowledge participants, varying whether they were given a goal at initial encoding (e.g., you will be traveling and need a camera that is easy to use), and whether they had a similar or different goal at each of two rounds of retrieval (Cowley & Mitchell, 2003). Learning by participants with low domain knowledge was guided by their initial goals and the fit between those goals and their first goal at recall, whereas high domain knowledge participants' learning was not influenced by the assigned goals at either encoding or recall.

The basic context of thinking about purchasing products is useful for studying learning. For example, consistent with work integrating research on categorical and quantitative inference learning (Juslin, Olsson & Olsson, 2003) and work discussing the importance of how categories are used in category learning (Markman & Ross, 2003), Eisenstein and Hutchinson (2006) examined inference learning tasks in a consumer behavior context. They provided participants with product features and asked them to make either a categorical or numeric judgment. Some estimated whether the product's price was above or below a specific price, and others estimated the product's price. Consistent with the importance of use, asking for categorical responses led to greater levels of learning for examples near the cutoff point (Eisenstein & Hutchinson, 2006). The implication from this work is that well-designed cutoffs can make learning efficient, but poor cutoffs can distort learning and undermine later category use.

Categories

Inference learning is just one instance of a much broader interest in categories within consumer behavior research. Consumers and marketers rely on categories to organize types of products and services (e.g., televisions, minivans, banking). They also rely on brand categories (e.g., McDonalds, Wal-Mart, Sony) to organize judgments of quality, availability, desirability, and other concerns. They also use social

categories to identify kinds of consumers (e.g., early adopters, coffee drinkers, penny-pinchers) that then guide how, for example, people design products, generate marketing campaigns, and choose retail outlets. This work points out intriguing possibilities for research on categories and concepts (see Rips et al., Chapter 11).

As an example, research on brand extensions raises interesting possibilities about category membership: If you have a choice of including an item in a category, forming a subordinate category or forming an independent category, what should you do? If Crest has a strong brand based on its toothpaste products, what might happen if the brand Crest was applied to other products, such as mouthwash, toothbrushes, sinks, dishwashing detergent, or glue? Would this generate better or worse appraisals of the other products, and does it matter what kinds of other products? Would the extension help or hurt the Crest brand as a whole, and would it help or hurt the main Crest toothpaste product on which the brand was built? These are pragmatic questions of interest to marketers, of course, but they are also interesting questions for the study of categories. They highlight the role of people's ability to generate and change categories and the effects of those changes on people's perceptions of the category and new and old category members. These questions generally do not arise when considering categories like robins or chairs. The consumer behavior context makes the roles of people's attitudes and social interaction more central to the study of categories.

There has now been at least 20 years of active experimental research on brand extensions in consumer behavior. This work has established that if the new product will be atypical of the category and not clearly understandable as a category member, then it is more likely to lower people's perceptions of the category as a whole (i.e., the brand; see Keller & Lehmann, 2006, for a review). In some cases, including the new product in the brand can even lower people's perceptions of the flagship product (i.e., central category member; John, Loken, & Joiner, 1998). The variety of products already included in the brand category is also an influence on people's willingness to accept a further product as an instance of the brand (Meyvis & Janiszewski, 2004).

Consumer behavior research on categories also highlights that categories can change over time. For example, the development of the minivan product category shows influence of both consumers and

producers (Rosa, Porac, Runser-Spanjol, & Saxon, 1999). Most basic research on categories, category membership, and category typicality has focused on stable properties of category members, rather than histories of the frequency of instantiation of items as category members (Barsalou, 1985). Examining magazine articles for frequency of instantiation—the proportion of product mentions that were made in reference to a product category, such as “the Honda Odyssey is a minivan that...”—provides a measure of typicality derived from real-world aggregate behavior (Rosa et al., 1999; see also Rosa, Judson, & Porac, 2005). This frequency of category instantiation measure predicted a product's longevity in the market—that is, the more often products were mentioned as being category members, the more likely they were to persist in the market. Frequency of instantiation may in practice be confounded with feature-based criteria for typicality. However, if categories themselves can change over time, and consequently what is central or ideal can change over time, then frequency of instantiation may be critical for establishing category typicality in practice. Given that people rely on a division of cognitive labor so that they do not need to become experts in all domains (e.g., Keil, Stein, Webb, Billings, & Rozenblit, 2008), and given that people are capable statistical learners (e.g., Saffran, Johnson, Aslin, & Newport, 1999), it is plausible that frequency of instantiation is a far greater influence on category learning than current psychological models imply.

Consumer behavior research on categories of services, such as cruises or massages, also extends basic research on event categories. For example, one of the key features of events is that they unfold over time. This allows event duration to be causally important. For example, people use a service's stated length as a heuristic guide to its value (Yeung & Soman, 2007). Furthermore, experiencing greater variety within an event led people to report immediately afterward that the event had a shorter duration than did those experiencing less variety (Ahn, Liu, & Soman, 2009). One to three days later, however, the people who experienced greater variety reported that the event had lasted longer. Thus, event duration can be a central feature of event categories.

Underlying much of this discussion of product and service categories is a close connection between category typicality and preferences. An examination of brands (e.g., Taco Bell, American Airlines) in eight product and service categories (e.g., types of restaurants, types of transportation) found strong

correlations (about .6) between a particular brand's typicality within a product category and positive attitudes toward that brand (Loiken & Ward, 1990). People's attitudes toward category members tend not to be discussed in basic research on concepts and categories. But consistent and strong relationships between typicality and attitudes are important to integrate theoretically and are certainly important practically for thinking about category effects.

Similarity and Analogy

Similarity and analogy research has also considered the joint action of cognition and affect (e.g., Thagard & Shelley, 2001), and consumer research offers new insights here as well. People use experiences as analogical bases that they extend by analogy to understand and develop emotional responses to new products or services serving as targets (Goode, Dahl, & Moreau, 2010). The strength of people's preferences for the base experiences together with the number of inferences drawn from base to target influence people's emotional responses to targets (Goode et al., 2010).

This work is part of a larger stream of research on analogy and similarity in consumer behavior. For example, companies use similarity and analogy to make their products comprehensible and appealing to consumers. Providing analogies appears to promote more focused knowledge transfer and inferences than assigning a product to a category, as categories license generating inferences about both surface features and underlying structure or relations, whereas analogies preferentially focus people on underlying structure (Gregan-Paxton & Moreau, 2003). Analogies are particularly useful for helping consumers understand very new products, rather than incrementally new products (Moreau, Markman, & Lehmann, 2001). Companies can also manipulate surface similarity by designing products to look like something with the same underlying function (Rindova & Petkova, 2007). For example, digital video recorders, such as TiVo, were designed to look like VCRs so that consumers spontaneously made appropriate comparisons. The broad implication, and one that will recur in other areas of business research on thinking and reasoning, is that an understanding of how people think can guide businesses to design more effective products, processes, and policies.

Creativity

Consumer behavior also provides useful contexts for examining creativity (see Smith & Ward,

Chapter 23). Creativity is important for understanding advertising. Some advertising is mainly concerned with raising consumer awareness of a product or brand, whereas other advertising is aimed at engaging consumers and increasing their attitudes toward and involvement with brands (Vakratsas & Ambler, 1999). It is for this latter group that creativity is consequential, because it can draw consumers' attention and consideration. One of the interesting outcomes of research on advertising is that it demonstrates that there are reliable ways to structure advertisements such that people will find them creative and effective (Goldenberg, Mazursky, & Solomon, 1999; Loewenstein, Raghunathan, & Heath, 2011; McQuarrie & Mick, 1996). An avenue for future cognitive research would be to unpack why these structures are effective and to identify additional structures (e.g., Loewenstein & Heath, 2009).

Creativity is also important in understanding how consumers use products. For example, one concern is what leads consumers to generate creative uses for consumer products, such as how one might generate one of those 1,001 uses for duct tape (Burroughs & Mick, 2004). Consumer research also provides novel contexts in which people can be asked to generate new and potentially creative kinds of category members, similar to Ward's (1994; Ward, Patterson, & Sifonis, 2004) drawing tasks. For example, online shopping allows consumers to design their own products (Dellaert & Stremersch, 2005), such as Nike enabling customers to design shoes.

This brief survey shows that thinking and reasoning research has important applications for understanding consumers and those who market to them. It also shows that the context of marketing to consumers spurs insights and reveals new phenomena to integrate back into basic research on decision making, learning, categories, similarity, and creativity, among other topics.

Management: Organizational Behavior

To have a product or service to sell takes work, of course. The field of organizational behavior aims to explain and facilitate employee, manager, and executive work, including, for example, decision making, conflict management, the generation and production of innovations, and group performance. Organizational behavior research generally relies on field surveys and experiments, but there is also qualitative research, research using archival sources, and a small amount of computational modeling work.

Ethics and Decision Making

Organizational behavior is the only behavioral research area in business in which decision-making research is not the dominant stream of work on thinking and reasoning (Moore & Flynn, 2008). Still, decision making has not been neglected. For example, decision making has been influential in organizational behavior research on negotiation (Bazerman, Curhan, Moore, & Valley, 2000). A distinctive contribution of organizational behavior research on decision making has been its examination of ethics (see Waldmann et al., Chapter 19). Ethics is related to practical concerns but also to a fundamental cognitive claim, which is that how people represent or frame a problem or decision influences the solution or choice people generate. An important consequence, as Tenbrunsel and Smith-Crowe (2008) note, is that if decision frames do not take ethics into consideration, then the resulting decisions may be unethical without people realizing it.

For example, one study engaged participants in a hypothetical dilemma: They could either cooperate with a group agreement at a high known cost to themselves or defect from the group agreement with a low known cost (Tenbrunsel & Messick, 1999). In a baseline condition, overall 55% stated they thought they were making an ethical decision, with the remainder framing it as a business (i.e., a narrowly framed cost/benefit) decision. In a second condition, in addition to the low known cost of defection, there was a chance of being caught for defecting and having to pay a small penalty. In this case just 18% used an ethical decision frame—the threat of a penalty shifted people's decision frames. The decision frame was consequential, as 91% of those with an ethical decision frame cooperated, whereas just 39% of those with a business decision frame cooperated. It is plausible that the habitual use of many representations relevant to the business domain, as well as other domains in life, implicitly encourages people to make choices that are, unintentionally, unethical. Therefore, a useful project for thinking and reasoning research would be to examine factors that foster flexibility in how people represent information—the cognitive skills behind ethical behavior could overlap considerably with those for creativity.

A further concern over ethical implications of behavioral decision making is that risk-seeking behavior can involve unethical behavior. For example, in one study, people were asked to complete an

anagram task, with an opportunity to score and pay themselves for their own performance (Schweitzer, Ordoñez, & Duouma, 2004). Participants were assigned different goals: challenging goals, readily achievable goals, or simply a request to do their best. They found that relative to the two easy goal conditions, participants with the challenging goals—who likely thought that their performance would not be sufficient to allow them to reach their goals—were twice as likely to cheat. Related work found that participants who had already completed a goal were more likely to take a risky gamble than accept a sure payoff, and the more they had surpassed their goal the more likely they were to take the risky gamble (Jeffrey, Onay, & Larrick, 2010). It is an open question whether in the aftermath of achieving a goal people will be not only more risk seeking but also more willing to be unethical.

To address individual-level biases toward unethical decision making, or more broadly, towards poor decision making, research need not look solely to individual-level solutions. Social and organizational contexts can mitigate some individual shortcomings (Heath, Larrick, & Klayman, 1998). For example, the guidelines for total quality management, a collection of practices aimed at improving the reliability of work processes, encourage people to gather complete information about problems, in contrast to individual tendencies to gather information selectively. A broad implication is that examining organizational routines and jobs can reveal specialized controls and tasks that cover for what people otherwise fail to consider or do. Another implication is that there are likely other cognitive limits that organizations are failing to address, awaiting people to draw attention to them and propose repairs.

Learning

Another central cognitive concern in organizational behavior is learning from experience, with a particular concern for learning from events that are often complex, ambiguous, and far from randomly sampled (Loewenstein & Thompson, 2006; March, 1994; Wood, 1986). For example, Denrell (2003; Denrell, Fang, & Levinthal, 2004; Denrell & March, 2001) used computational modeling to show that organizational settings produced a host of biased samples as successful managers and organizations tend to stay and grow in prominence, whereas unsuccessful ones tend to leave or fold. The result is effectively learning based on sampling on the dependent measure.

Another approach to learning from experience is to examine counterfactual reasoning. For example, one study examined airplane accidents (Morris & Moore, 2000). The National Transportation and Safety Board maintains records of airplane accidents and near-accidents. These include narrative descriptions of accidents, which can be coded for the presence of counterfactuals and the presence of specific lessons for the future. Pilots' narratives with upward, self-directed counterfactuals (e.g., If only I had done X, the accident would not have happened) were most likely to exhibit lessons for the future, in contrast to narratives with downward counterfactuals (e.g., If I had not done X...), narratives with other-directed counterfactuals (e.g., If only air traffic control had said...), and narratives with no counterfactuals (Morris & Moore, 2000). Two follow-up experiments tested the effects of counterfactual reasoning on learning from experience used a flight simulator game. Replicating the archival study of actual pilots, participants who spontaneously generated upward, self-directed counterfactuals were most likely to write specific lessons learned. Participants asked to generate upward or downward self-directed counterfactuals after an initial flight simulator session revealed that those asked to make upward counterfactuals showed greater improvement on a subsequent flight simulator session than those asked to make downward counterfactuals (Morris & Moore, 2000). These studies also found that people generated more counterfactuals the more severe the accident, and the more personally accountable (as opposed to organizationally accountable) the pilot. The implication is that counterfactual reasoning is a form of self-explanation that people generate spontaneously in response to events that can yield learning.

Categories

Organizational behavior has a long-standing interest in the products of learning. For example, there is research on novice and experienced entrepreneurs' understandings of worthwhile and non-worthwhile opportunities to start a new business (Baron & Ensley, 2006), managers' understandings of threats and opportunities (Jackson & Dutton, 1988), and participants' reactions to tasks framed as work and play (Glynn, 1994). In a characteristic early study along these lines, Lord, Foti, and De Vader (1984) examined the category "leader," using feature listing and ratings tasks to assess what features indicate the category. They showed that

participants rated a feature's prototypicality faster the more typical it was. They also showed that a vignette about a leader managing a project that used typical features for the leader, compared to one using atypical features, yielded ratings indicating that participants expected the leader to behave in more typical ways, and considered the leader more responsible and accountable for the project.

In a series of studies on firm competitors, Porac and colleagues (Porac & Thomas, 1994; Porac, Thomas, & Badenfuller, 1989; Porac, Thomas, Wilson, Paton, & Kanfer, 1995; Porac, Wade, & Pollock, 1999) examined managers' category taxonomies as an influence on which firms a manager at a given company would say were competitors. Rather than assuming managers are rational agents examining every firm's resources and dependencies to determine their competitors, this work suggests that managers use categories of firms to simplify and guide their attention. For example, one study examined a category taxonomy derived from the superordinate category "retailer" (Porac & Thomas, 1994). Store managers noted typical subordinates such as groceries and drug stores and atypical subordinates such as travel agencies and animal grooming shops. Store managers rated stores from the same subordinate categories as their own stores as far more likely to be competitive threats than those from different subordinate categories. Store managers also tended to perceive typical retailers to be more of a competitive threat than atypical retailers. Thus, category membership and typicality influenced managers' perceptions of the firms with whom they were competing.

A further study surveyed managers from the Scottish knitwear industry (Porac et al., 1995). They found that five prominent dimensions (firm size, location, knitting method, construction method, and product style) accounted for how managers categorized firms in the industry. Managers largely perceived their competitors to be other firms within the same category as their own firm. Furthermore, the more typical a firm was of its category, the more likely it was to be rated a competitor by others in its category. These studies indicate that category taxonomies and category typicality are tools managers use to simplify the task of who to monitor and compete against. They also suggest that there is social consensus about categories.

Organizational research, due to concerns over power and ethics, also highlights that people sometimes manipulate category membership. For

example, public companies have to report which firms they used as a basis for comparison when deciding on the compensation package for their CEOs. Usually the reports cite other firms from the same industry. However, managers include comparison firms from outside their industry in greater numbers when, for example, firm performance is low or when their industry performs well (Porac, Wade, & Pollock, 1999). The implication is that managers are changing category membership when it serves the purpose of rationalizing higher compensation for their CEOs. Thus, firm categories are useful and cannot simply be ignored, but they can be manipulated.

Another area of research examines culture's influence on categories. For example, Keller and Loewenstein (2011) used a cultural consensus model analysis, following Atran, Medin, and Ross (2005), to examine how people think about a complex category: cooperation. The study examined similarities and differences in how people in the United States and China understand 17 dimensions related to the category of cooperation. Overall, there was substantial consensus across both nations about what does and does not indicate cooperation. Importantly, there were high levels of consensus both for settings that indicate cooperation (e.g., having aligned goals, being friends with the people in one's group) and for actions that indicate cooperation (e.g., putting in effort on group tasks, sharing knowledge with others). This makes the category of cooperation complex and not easily defined, but it also makes the category of cooperation highly useful. Because the category spans settings and actions, it performs a valuable function by linking perception and action. A broad implication might be that internal category structure might be complex so as to increase the efficiency of category use.

The study of cooperation also found a striking cultural difference in category membership. About two-thirds of the Chinese participants (and a third of U.S. participants) felt that competing with others within one's group and trying to outperform them was considered cooperative, and that not trying to compete within one's group was considered noncooperative. Additional research (Keller, Loewenstein, & Yan, 2010) shows that people in both cultures believe that cooperation and competition are opposites. The difference seems to be that people in China are more likely to show a predilection for dialectics (Peng, Spencer-Rodgers, & Nian, 2006),

or an interest in integrating opposites rather than keeping them separate. Dialecticism mediates the link between culture and the categorization of competitive behaviors within a group as cooperation. Broad cultural tendencies in how to reason about categories can influence category content and the relationship among categories.

The cultural differences and uses of categories highlight social bases for category formation that are generally not considered by basic psychological research. Most such research assumes that what makes something a member of a category is the intrinsic properties of the category members themselves (e.g., what features it has). Additional research examines the particular goals people have in using the categories (Ratneshwar, Barsalou, Pechmann, & Moore, 2001), and the role of the category member in a larger event or system of relations (relational and role-based categories; Gentner & Kurtz, 2005; Markman & Stilwell, 2001). Yet although it is only rarely described this way, category membership in most basic research is dictated by experimenter fiat, and participants usually manage to learn the categories. Categorization research needs a means for separating out the role of categorization based on features from categorization based on social or cultural forces.

Creativity

Research on creativity is also an important organizational behavior topic, as innovation is a common reason for the creation for new businesses and the success of existing businesses. A central finding from this work is that bringing knowledge from a variety of domains or contexts together is important for creativity (George, 2007). For example, a case study found that designers at IDEO are effective because they bring together people with experience working on different kinds of products, enabling them to draw effective analogies (Hargadon & Sutton, 1997). This is similar to Dunbar's (1995) case study research on microbiology lab groups (see Dunbar & Klahr, Chapter 35). An archival study of comic book prices found that on average, the more genres in which a comic book's authors had experience, the higher the market value of the comic book (Taylor & Greve, 2006). There was value for both individuals and teams of authors to have had varied domain experience, and signs that individuals who themselves had a variety of prior experience had the strongest relationship between variety and creativity.

A further study of creativity examined the researchers at an applied research institute, their working relationships, and the creativity of their work. The greater the variety of work experience represented in a researcher's relationships, the more likely they were to have their work rated as creative (Perry-Smith, 2006). For these researchers, the variety of work experience that was important was not those with whom one worked closely (strong ties, in social network parlance), but those with whom one was acquainted (weak ties). These illustrative studies of creativity using a variety of methods rarely used in basic research suggest that there is robust support for the claim that the variety of knowledge one can bring to bear is important for producing creative work. It also suggests there is still unexplained variation as to when it is most beneficial for that variety of knowledge to be mastered by one person, held within a tight working group, or accessed through brief interactions.

Group Cognition

The field of organizational behavior also has a strong interest in how groups and organizations learn, remember, innovate, and perform. Thus, cognition research in organizations is frequently at the level of social aggregates (e.g., Walsh & Ungson, 1991). These are not claims about group minds. Rather, the interest is in how people work together, socially distributing information-processing tasks and coordinating problem-solving activity. For example, people working closely together learn to generate divisions of cognitive labor, allocating different kinds of information about their tasks to different members to encode, remember, and retrieve from each other when needed (Faraj & Sproull, 2000; Lewis, Lange, & Gillis, 2005). At the level of the organization, research on learning curves, for example, has demonstrated that manufacturing error rates decrease and efficiency increases with production experience, echoing individual-level learning curve research (see Argote, 1999, for a review). Thus, there are literatures within organizational behavior on how large-scale cognitive tasks are socially distributed and the cognitive challenges and opportunities involved. The implication for research on thinking and reasoning is that analyzing individual thinking and reasoning may be misleading if the activities under study are collaborative, and that to understand the collaborative cognitive work that consumes so many people's working lives, studies of individual thinking and reasoning can offer only a partial account.

Taken together, organizational behavior research on thinking and reasoning highlights the value of studying cognition in context. For example, the categories research shows why categories may be so complex: social and cultural influences can dictate whether something is or is not a member of a category regardless of what a given individual might believe, and category membership can be complex if it serves to simplify category use. Furthermore, because people are using categories, making decisions, learning, and being creative in the context of working with others, then the thinking and reasoning they do may be qualitatively different due to those interactions and due to the distribution of thinking and reasoning across members of a team or organization.

Operations: Management Science

Business activity often now involves considerable technical complexity in buying, developing, manufacturing, and selling products. It also involves gathering, generating, processing, and distributing large quantities of data. Management science emphasizes the role of analytic and information technology support for guiding organizational action and decision making. It has something of an engineering culture and emphasizes mathematical and computational modeling to guide potentially complex, practical and large-scale concerns, such as scheduling transportation, maximizing factory production yields, and minimizing risk. Management science research links to thinking and reasoning research mainly in considering the use and flow of information, but there is also a stream that considers the pragmatic, robust implementation of routines for effective performance despite the involvement of fallible human performers. Management science research often involves formal and computational modeling but also includes field research, archival research, and experimental research.

Cognitive Support Systems

One concern in management science is to design support systems to process information in the service of improving decision making. Consistent with Payne and others' research on multiple goals in decision making noted earlier, Todd and Benbasat (1992) found that people tend to use decision support systems to reduce the effort involved in decision making rather than to improve the quality of their decision making. Furthermore, if people perceive that decision systems will require effort to use

or will function in a rigid way, they are less likely to want to use them (Wang & Benbasat, 2009). Novices seem particularly strongly guided by concerns about ease of use, whereas experienced users are more evenly guided by ease of use and ability to control the functioning of the decision support system (Taylor & Todd, 1995). Why and how people choose to use decision support systems appear to be consistent with research on unaided decision making, and the interaction between the two areas should lead to more effective support systems.

The conclusions from decision support systems are consistent with broader examinations of why people choose to use any kind of information technology (Venkatesh, Morris, Davis & Davis, 2003). There are thus more areas to learn about by examining information technology use than decision making alone. For example, different communication media serve different communication needs, so examining which media fit which kinds of tasks could lead to improved performance (Te'eni, 2001). New information technologies provide new opportunities for creating fits, and how people choose among available communication media can therefore be informative (Watson-Manheim & Bélanger, 2007). As a further example, information systems frequently generate classification systems, and research on human categorization is beginning to guide the design of such systems (Parsons & Wand, 2008). The general point is that analyzing the features of artifacts designed to support cognition provides avenues for understanding unaided cognitive processing, and management science research has yielded a wealth of data on this issue.

Information Search

Management science also contributes to thinking and reasoning research by addressing neglected topics. For example, most basic research provides information to participants rather than examining how people search for information or examining when people decide to stop searching for information. The widespread use of information technology, company databases, and the Internet makes the search for information salient, and satisficing (Simon, 1955) provides a reason to think about people's rules for stopping their search for information. One study examined five stopping rules: finding information about a single issue; finding all the information on one's initial list; noting that one's representation is no longer changing; noting that the incremental gain of each new piece of

information is consistently small; and noting that the amount of information gathered reached a threshold (Browne, Pitts, & Wetherbe, 2007). Different stopping rules seemed to fit with different kinds of tasks. For example, the list stopping rule was commonly used by people engaged in product and job search tasks, whereas the threshold stopping rule was commonly used by people engaged in a map search task. This line of research could be broadly expanded and doing so could contribute substantially to understanding real-world problem solving and decision making.

Learning

Management science research also has an interest in how people learn to act in complex and dynamic systems. For example, Sterman (1989a) examined people's performance on a dynamic inventory distribution task, the classic "beer game." In the task, four people play roles in a typical supply chain: A beer manufacturer supplies a beer distributor, who supplies a beer wholesaler, who supplies a beer retailer. As customers purchase beer from retailers, retailers need to request new supplies, and those requests ripple up the supply chain, with lags at each step. People learn poorly from the delayed feedback built into the task (Sterman, 1989a). Dynamic systems modeling of the tasks facilitated pinpointing the locus of participants' errors. Tracings of the data implied that participants were using an anchoring and adjustment strategy in their largely futile attempts to keep their supplies constant. Open-ended participant responses afterward confirmed that the impression from the model parameters fit people's subjective impressions. In additional research (Sterman, 1989b), participants in a single-player dynamic game showed similar patterns.

Interaction with a dynamically changing environment and delayed feedback is a common and often unavoidable aspect of action in the world, and it clearly presents strong challenges for learning (Rahmandad, Repenning, & Sterman, 2009). It is less often a feature of cognitive psychological research, but it is tractable. For example, part of people's difficulty with learning from delayed feedback is a failure to keep track of prior actions (Gibson, 2000). Modeling and experimentation showed that people's learning from delayed feedback improved if the system displayed their history of prior actions. Similar issues regarding delayed feedback arise in computer-based tutoring systems (e.g., Anderson, Corbett, Koedinger, & Pelletier, 1995), but with an added

concern with task dynamics. Delayed feedback is not the only aspect of dynamic systems that is challenging conceptually. Even in very simple situations, people have difficulty understanding accumulation (Cronin, Gonzalez, & Sterman, 2009).

Groups

A final feature of management science research to note is that there is a concern for how groups and organizations work together using information technology to accomplish tasks. For example, large-scale projects, such as building an airplane or writing and revising the Linux operating system, raise the issue of how to design the project so that multiple people can work on the project. If working on one part of the project depends on what happens on another part of the project, this makes doing the work more complicated, error prone, and costly. MacCormack, Rusnak, and Baldwin (2006) analyzed open source software design projects, using design structure matrices to capture the interdependencies in the source code. This method allowed them to analyze and track changes in the overall pattern of dependencies among parts of the project, providing useful insights into quantifying the design of projects to estimate the likelihood of effectively distributing the work, or what they call the “architecture for participation.” It is plausible that reducing dependencies among parts of a problem would also ease individual problem solving.

Management science research, by highlighting roles for technology in thinking and reasoning and distributing thinking across individuals working together, raises similar issues to discussions of situated cognition in the 1990s. Most real-world thinking and reasoning involves groups of people generating and using artifacts to accomplish tasks, and hence studying individuals in isolation is likely insufficient to explain the thinking and reasoning of people at work. With the proliferation of information technology, its ready accessibility in the form of mobile phones and tablets, and the geographic dispersion of work, the work in management science to understand how groups of people work and use technology to assist their work provides further support for the practical value of that research agenda and provides new reasons to extend it.

Finance: Behavioral Finance

Part of the practice of business—and to some, perhaps too much of the practice of business—is concerned with financial markets. Behavioral

finance aims to explain actors’ decisions in financial markets, providing an alternative to financial economics models that assume markets are aggregates of the behaviors of rational agents. As psychological research has called into question the assumption that individual agents act rationally, behavioral finance researchers have begun to test for effects of biased actors on financial activity, and to formulate new models for understanding individual investor behavior and aggregate market performance (Barberis & Thaler, 2003; Subrahmanyam, 2007). This research involves core issues in finance, such as whether stock prices are accurate representations of the aggregate information about the prices of firms. Behavioral finance research consists primarily of experimental research, archival research to examine individual and aggregate performance of financial actors, and mathematical modeling.

Judgment and Decision Making

The most proximal cognitive concern in behavioral finance research is work examining the behavior of individual investors. This is a decision-making context about which one can sometimes obtain excellent real-world behavioral data, such as transaction data from individual retirement accounts. It is also an important practical area to study. For many reasons, including the shift in retirement plans away from providing a defined benefit or pension to providing a defined contribution to an investment portfolio that individuals themselves manage, more people than ever are making investment decisions.

Unfortunately, people often appear to make poor financial investment decisions. There is both experimental evidence and evidence from investment records that employees tend to allocate their own retirement investment funds based on the choice sets provided by their company retirement plans (Benartzi & Thaler, 2001). The array of funds offered by companies play an important signaling role for what investments people perceive to be normative. For example, in one experiment, participants tended to allocate money evenly between two mutual funds, regardless of whether the two funds’ holdings were stocks and bonds, stocks and a mixture of stocks and bonds, or bonds and a mixture of stocks and bonds (Benartzi & Thaler, 2001). This pattern matches a tendency to seek variety from options that are presented simultaneously that is also found in consumer behavior decision-making research (e.g., people’s selections from arrays of snack foods; Simonson, 1990). Choices

made serially show less variety-seeking behavior, as people repeatedly choose a favored item. Accordingly, Benartzi and Thaler's (2001) conclusion was that people make major financial decisions based on the same heuristic that guides their selection of candy bars. Behavioral decision-making tendencies appear robust, for good and ill.

A further general point is that people seem to make decisions as if each decision was separate from others (choice bracketing; Read, Loewenstein, & Rabin, 1999). Standard finance models assume that people's choices regarding an investment decision should be linked to one's other investments, one's home mortgage, and so forth. Instead, it appears that people tend to treat these decisions as if they were independent. This tendency relates to the notion that people make choices with respect to a reference point, and hence people are sensitive to changes in utility, not absolute or total utility (Tversky & Kahneman, 1974). The reference point is often used to highlight the difference between how people perceive gains and losses, and the loss aversion effect, or the tendency to avoid losses more strongly than to seek comparably sized gains. But just as important as loss aversion is the notion that the reference point is generated separately for each decision—decisions are narrowly framed or bracketed (Barberis, Huang, & Thaler, 2006). Without this assumption, prospect theory and related models could not predict the basic loss aversion effect. Furthermore, if people think about financial decisions separately rather than in terms of their influence on their total financial portfolio and are risk averse, this could help explain the historical reluctance of individuals to invest in the stock market—in 1984, half the households with \$100,000 in liquid assets did not own stock (Barberis et al., 2006).

Vivid information appears to have large-scale effects due to its influence on decision making. Following Griffin and Tversky's (1992) suggestion that people overemphasize vivid evidence and underemphasize the credibility of the source of that evidence, Sorescu and Subrahmanyam (2006) tested the implications of buying or selling stocks based on large or small changes in recommendations (e.g., strong sell, sell, hold, buy, strong buy) by high- and low-credibility analysts. Examining stock prices over several days, all recommendations, on average, yielded gains (positive abnormal returns). However, if the evidence was vivid (i.e., a large change in recommendation) and made by a less credible source (analysts with little experience or who work for

lower prestige firms), then over the longer term, the recommendations led to losses. In contrast, if the evidence was not vivid (small change in recommendation) or was made by a higher credibility source (analysts with years of professional experience or who work for the highest prestige firms), then the recommendations yielded gains over the longer term (Sorescu & Subrahmanyam, 2006). The implication is that stock investors appear to overreact to low quality, vivid evidence, consistent with claims from lab research.

Behavioral finance generates some striking evidence that irrelevant or transient information can be consequential. For example, stock market returns tend to be positive on sunny days and mixed on cloudy days (Hirshleifer & Shumway, 2003; Saunders, 1993). Similarly, the outcomes of national sporting events predict national stock market performance (Edmans, Garcia, & Norli, 2007). Even simple association effects matter. Company names that capitalize on cultural trends show stock market advantages. Companies adding dot-com names during the technology boom of the 1990s generated sharp gains after the announcement of their name change (Cooper, Dimitrov, & Rau, 2001). The gains did not fade immediately afterward but appeared to persist. Furthermore, companies dropping dot-com names during the technology crash also generated sharp gains (Cooper, Khorana, Osobov, Patel, & Rau, 2005). Some clever companies both added and removed dot-com names, and they tended to generate stock price gains both times. Similarly, mutual fund name changes attract investors (and hence money), but have no influence on fund performance (Cooper, Gulen, & Rao, 2005). The emotion-cognition link (e.g., Clore & Huntsinger, 2007) has broad social consequences.

Prediction markets show another interaction between individual-level thinking and macro behavior. In prediction markets, people wager real or virtual money on the outcomes of future events, such as the revenue of Hollywood movie openings, the level of U.S. unemployment, the winners of political elections, and the outcomes of geopolitical events (Wolfers & Zitzewitz, 2004). The logic is that the opportunity to make money provides an incentive to find, aggregate, and weigh information to form judgments. However, lab studies show that because people sometimes make optimistic bets, other market participants observing the bets tend to change their beliefs to believe the optimistic outcome is more likely (Seybert & Bloomfield, 2009).

Thus, individual beliefs interacting with collective behavior can make beliefs more and less accurate, and work on prediction markets is aimed at understanding when each occurs.

Learning and Expertise

Another main finding in behavioral finance is that investors who are confident in their own abilities make more trades (Graham, Harvey, & Huang, 2009). This suggests the importance not only of examining how people make a given decision but also the reasons why they are likely to make a decision or take an action in the first place (Heath & Tversky, 1991). For example, in a series of studies of archival records of individual brokerage accounts, Odean (1998, 1999; Barber & Odean, 2000) generated evidence consistent with the possibility that overconfidence leads investors to buy and sell stocks. Odean and colleagues found that those who trade more earn less, and that the stocks they sold perform better in the coming year than the stocks they purchased.

Investment experience does not necessarily alter cognitive tendencies. For example, records of bond futures traders' personal trading accounts show that those who lost money during the first half of the trading day made more and riskier trades during the second half of the trading day, consistent with loss-aversion-induced, risk-seeking behavior (Coval & Shumway, 2005). It is as if professional traders tally their accounts at the end of each day and wish to avoid finishing the day with a loss. Of interest, the assumptions that traders are loss averse and tally their accounts at the end of each day are key to a model of investor behavior to explain why stocks have, historically, yielded higher returns than bonds (the equity premium puzzle; Benartzi & Thaler, 1995). Thus, research on decision making appears to be a useful guide to at least some forms of important investment activity by both novices and experts; furthermore, cognitive research may provide useful components for models of aggregate behavior (cf., Goldstone, Roberts, & Gureckis, 2008).

Overall, behavioral finance research shows that numerous individual-level cognitive tendencies have large personal consequences and are useful predictors of aggregate behavior. It is heartening to see basic research hold up in real-world settings under high stakes. It is also a spur to search for more comprehensive models that incorporate additional aspects of thinking and reasoning, and to craft interventions to support people's thinking

and reasoning when they are making large financial decisions.

Accounting: Behavioral Accounting

Businesses aim to make money, which sounds simple but can be extremely complex to assess, let alone bring about. Behavioral accounting research examines how people generate, use, audit, and regulate an organization's quantitative economic information (Libby, Bloomfield, & Nelson, 2002; Sprinkle & Williamson, 2006). Accounting research is concerned with how such quantitative information is used within organizations (the domain of managerial accounting), such as for budgeting or compensation systems to pay employees. Accounting research is also concerned with how quantitative economic information is used by organizations to communicate to lenders, investors, and regulators (the domain of financial accounting), such as for financial statements the Securities and Exchange Commission requires of public companies. In all cases, there is a concern with generating accurate quantitative information for guiding decision making and behavior.

Expertise

A major tenet of how financial markets function is that markets accurately integrate information into prices. The accounting literature shows that the analysts whose judgments guide the buying and selling of stocks and other concerns traded in financial markets form systematically inaccurate assessments. One reason is that they rely too heavily on domain categories. For example, companies can use accounting categories to frame their actions in ways that influence financial analysts' stock price judgments (Hopkins, 1996). To establish baselines, Hopkins showed that an announcement that a firm acquired financing through a loan led analysts to predict no change in stock prices, whereas announcing financing through the use of additional stock led analysts to predict a 4% drop in stock prices. Hopkins (1996) provided other participants with an announcement of mandatorily redeemable preferred stock that was framed as either being like a loan or like additional stock, which yielded predictions comparable to the effects to the unambiguous frames. Thus, rather than generating the same predictions from identical financial terms, analysts with an average of 10 years of professional experience instead appeared to rely on category-based induction to guide their judgments.

As a further example, another study contrasted two forms of presenting information about firm actions on analyst forecasts (Sedor, 2002). One form was a list of facts. The other form was a narrative linking the facts into a coherent explanation. Analysts with an average of 8 years of professional experience forecasted higher earnings from the narrative form, consistent with a range of research on the value of information in story form (e.g., Pennington & Hastie, 1986) and making relations explicit (e.g., Gentner & Toupin, 1986). The implication from this work and other behavioral accounting research (Bonner, 2008) is that cognitive information-processing models make predictions that fit the judgment and decision-making performance of professional accountants and analysts. Furthermore, the research suggests that companies are taking advantage of these information-processing tendencies to generate favorable market reactions.

Another issue raised in behavioral accounting research is predicting effective knowledge transfer. Most cognitive science research examines learning and knowledge transfer in settled domains of math and science, such as probability judgments (e.g., Ross, 1987). In this way, researchers can establish correct answers. In domains such as accounting (and many other areas of business, and, for that matter, life), it is far more challenging to establish that knowledge transfer is useful rather than misleading. Cognitive science research on knowledge transfer (e.g., Gick & Holyoak, 1980; Loewenstein, Thompson, & Gentner, 1999; Singley & Anderson, 1989), combined with domain task analysis (e.g., Bonner & Pennington, 1991, provided a general task analysis for the auditing process) could therefore help to identify opportunities for effective knowledge transfer. The result could produce substantial savings in training and benefits for task performance.

For example, auditors with experience in the financial services industry performed as well at assessing the potential for a manufacturing firm to go bankrupt as auditors from the manufacturing industry (Thibodeau, 2003). They were also better than analysts in the manufacturing industry at assessing bankruptcy potential for firms in the gambling industry. Thibodeau (2003) measured auditors' knowledge of the core accounting tasks for assessing firm financial conditions, future cash flows, and payment histories, and showed that this knowledge mediated auditors' assessments. Thus, cognitive research can facilitate understanding the skills involved in task performance, which in turn

can guide the search for opportunities for effective knowledge transfer.

Numerical Cognition

Given the strong focus on handling quantitative information in accounting, behavioral accounting research has examined managers' and accountants' memory and recall of numeric accounting data (Kida & Smith, 1995; Libby, Tan, & Hunton, 2006). For example, in one study, researchers presented managers with accounting information for a specific firm as well as prior year figures and industry averages and then, an hour later, gave the managers a recall test (Kida et al., 1998). Managers remembered the affective tone of the accounting information most often, less frequently remembered the relative standing of the information (e.g., higher or lower than the prior year), and still less frequently remembered the approximate (let alone the exact) numbers.

In a further study, researchers presented managers with information about two sets of companies, separated by an hour's delay. After a further hour, they asked for the firm in which managers would be most interested in investing, and they found that a firm from the first set, which clearly dominated the set, was chosen more than a firm from the second set that did not clearly dominate its set but was clearly objectively superior (Kida et al., 1998). The implication is that managers' memory for numbers may preserve affective and qualitative aspects more robustly than the exact quantities, and that as a result making decisions from one's memory of numeric data can readily become distorted. This finding relates to consumer behavior research discussed earlier (Dickson & Sawyer, 1990) that showed consumers poorly recall exact prices.

This very brief review of behavioral accounting research provides an indication of the degree to which limitations to human thinking and reasoning, over and above decision biases, has important practical consequences for the practice of business. Behavioral accounting research might be very useful to link to research on thinking and reasoning in the law (see Spellman & Schauer, Chapter 36), which, despite having less of a numerical component, is also heavily concerned with understanding people's thinking when it is guided by a complex system of socially generated rules.

Conclusions and Future Directions

The practice of business has wide scope, making many aspects of thinking and reasoning important

for understanding this important applied area. Each area of business research examines thinking and reasoning involved in tasks in a major aspect of the practice of business. These are important practical concerns, so it is noteworthy for basic researchers to know that their work is being used to understand and improve behavior in these domains, and that there are applications of their work that they may not have realized.

There are many opportunities for exchange between basic and applied business researchers studying thinking and reasoning. Translating and applying basic research to understand a particular context raises questions, and studying particular applied contexts generates new phenomena and ideas. One goal of this chapter has been to highlight examples and so foster such work, in part by helping to develop mutual understanding despite differences in vocabularies (Clark, 1996). Thus, perhaps the simplest use of this overview is to provide starting points for scholars, journals, and topic keywords to examine for relevant research. It is easy enough to miss the fact, for example, that consumer behavior research on brands, organizational behavior research on competitors or organizational identities, and behavioral finance research on portfolio allocations, is all research on categories. These and other lines of business research highlight aspects of cognition that are less often considered in standard psychological research, such as (to stick with the topic of categorization) the potential for categories to change (e.g., membership changes, typicality changes), the potential for category membership to be a choice, or the influence of social and cultural factors on category membership.

An overarching theme from most areas of business academia is linking individual- and social-level cognition and behavior, whether by social one means groups, organizations, industries, or societies. Research on thinking and reasoning has the potential to help understand large-scale patterns of behavior (e.g., Goldstone et al., 2008; Koonce & Mercer, 2005; Loewenstein & Heath, 2009). Furthermore, effects at those social levels in turn influence individual cognition, such as what people attend to, learn, and think about (e.g., DiMaggio, 1997; Douglas, 1986; Ocasio, 1997). For example, due to the limitations of human memory, record-keeping was arguably critical for the development of advanced economies (Basu, Kirk, & Waymire, 2009). Archival research shows associations between the use of writing systems and the size of communities, the use of credit, and the

frequency of interactions between strangers (Basu et al., 2009). Whether it is to study culture, markets, organizations, industries, professions, or other macro concerns, research on thinking and reasoning can offer key insights.

A linked theme is that business academia makes far more use of survey and archival research methods than does basic research on thinking and reasoning. These methodological approaches offer the potential to examine sophisticated questions outside of the lab and thereby provide evidence of real-world consequences and open up possibilities for guiding organizational action and public policy. The broader implication is that a better understanding of cognition across levels of analysis would serve to enrich research on both individual cognition and the practice of business. It might also be profitable.

Acknowledgments

I wish to thank Andres Almazan, Susan Broniarczyk, Keith Holyoak, Steve Kachelmier, Lisa Koonce, Bobbie Spellman, and Michael Williamson for guidance and insights, and support from the Kelleher Center and IC², both of The University of Texas at Austin.

References

- Ahn, H.-K., Liu, M. W., & Soman, D. (2009). Memory markers: How consumers recall the duration of experiences. *Journal of Consumer Psychology*, 19, 508–516.
- Alba, J. W., & Hutchinson, J. W. (1987). Dimensions of consumer expertise. *Journal of Consumer Research*, 13, 411–454.
- Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, 4(2), 167–207.
- Argote, L. (1999). *Organizational learning: Creating, retaining and transferring knowledge*. Boston, MA: Kluwer.
- Atran, S., Medin, D. L., & Ross, N. O. (2005). The cultural mind: Environmental decision making and cultural modeling within and across populations. *Psychological Review*, 112(4), 744–776.
- Barber, B. M., & Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *Journal of Finance*, 55, 773–806.
- Barberis, N., Huang, M., & Thaler, R. H. (2006). Individual preferences, monetary gambles, and stock market participation: A case for narrow framing. *American Economic Review*, 96(4), 1069–1090.
- Barberis, N., & Thaler, R. H. (2003). A survey of behavioral finance. In G. M. Constantinides, M. Harris, & R. Stulz (Eds.), *Handbook of economics of finance* (pp. 1051–1121). Amsterdam, Netherlands: Elsevier.
- Baron, R. A., & Ensley, M. D. (2006). Opportunity recognition as the detection of meaningful patterns: Evidence from comparisons of novice and experienced entrepreneurs. *Management Science*, 52(9), 1331–1344.
- Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in

- categories. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 31(4), 629–654.
- Basu, S., Kirk, M., & Waymire, G. (2009). Memory, transactions records, and *The Wealth of Nations*. *Accounting, Organizations and Society*, 34, 895–917.
- Bazerman, M. H., Curhan, J. R., Moore, D. A., & Valley, K. L. (2000). Negotiation. *Annual Review of Psychology*, 51, 279–314.
- Benartzi, S., & Thaler, R. H. (2001). Naïve diversification strategies in defined contribution savings plans. *American Economic Review*, 91, 79–98.
- Benartzi, S., & Thaler, R. E. (1995). Myopic loss aversion and the equity premium puzzle. *Quarterly Journal of Economics*, 110, 73–92.
- Bettman, J. R., Luce, M. F., & Payne, J. W. (2008). Consumer decision making: A choice goals framework. In C. Haugvedt, P. Kerr, & F. Kardes (Eds.), *Handbook of consumer psychology* (pp. 589–610). Mahwah, NJ: Erlbaum.
- Bonner, S. E. (2008). *Judgment and decision making in accounting*. Upper Saddle River, NJ: Pearson.
- Bonner, S. E., & Pennington, N. (1991). Cognitive processes and knowledge as determinants of auditor expertise. *Journal of Accounting Literature*, 10, 1–50.
- Broniarczyk, S. M. (2008). Product assortment. In C. Haugvedt, P. Kerr, & F. Kardes (Eds.), *Handbook of consumer psychology* (pp. 755–779). Mahwah, NJ: Erlbaum.
- Broniarczyk, S. M., Hoyer, W. D., & McAlister, L. (1998). Consumers' perceptions of the assortment offered in a grocery category: The impact of item reduction. *Journal of Marketing Research*, 35, 166–176.
- Browne, G. J., Pitts, M. G., & Weatherbe, J. C. (2007). Cognitive stopping rules for terminating information search in online tasks. *MIS Quarterly*, 31(1), 89–104.
- Burroughs, J. E., & Mick, D. G. (2004). Exploring antecedents and consequences of consumer creativity in a problem-solving context. *Journal of Consumer Research*, 31, 402–411.
- Chernev, A. (2003). When more is less and less is more: The role of ideal point availability and assortment in consumer choice. *Journal of Consumer Research*, 30, 170–183.
- Chernev, A. (forthcoming). Product assortment and consumer choice: An interdisciplinary review. *Review of Marketing Research*.
- Chowdhury, T. G., Ratneshwar, S., & Desai, K. K. (2009). The role of exploratory buying behavior tendencies in choices made for others. *Journal of Consumer Psychology*, 19, 517–525.
- Clark, H. H. (1996). *Using language*. New York: Cambridge University Press.
- Clore, G. L., & Huntsinger, J. R. (2007). How emotions inform judgment and regulate thought. *Trends in Cognitive Science*, 11, 393–399.
- Cooper, M. J., Dimitrov, O., & Rau, R. (2001). A rose.com by any other name. *Journal of Finance*, 56(6), 2371–2388.
- Cooper, M. J., Gulen, H., & Rau, P. R. (2005). Changing names with style: Mutual fund name changes and their effects on fund flows. *Journal of Finance*, 60(6), 2825–2858.
- Cooper, M. J., Khorana, A., Osobov, I., Patel, A., & Rau, P. R. (2005). Managerial actions in response to a market downturn: valuation effects of name changes in the dot.com decline. *Journal of Corporate Finance*, 11, 319–335.
- Coval, J. D., & Shumway, T. (2005). Do behavioral biases affect prices? *Journal of Finance*, 40(1), 1–34.
- Cowley, E., & Mitchell, A. A. (2003). The moderating effect of product knowledge on the learning and organization of product knowledge. *Journal of Consumer Research*, 30, 443–454.
- Cronin, M. A., Gonzalez, C., & Sterman, J. D. (2009). Why don't well-educated adults understand accumulation? A challenge to researchers, educators, and citizens. *Organizational Behavior and Human Decision Processes*, 108, 116–130.
- Dellaert, B. G. C., & Stremersch, S. (2005). Marketing mass-customized products: Striking a balance between utility and complexity. *Journal of Marketing Research*, 42, 219–227.
- Denrell, J. (2003). Vicarious learning, undersampling of failure, and the myths of management. *Organization Science*, 14(3), 227–243.
- Denrell, J., Fang, C., & Levinthal, D. A. (2004). From t-mazes to labyrinths: Learning from model-based feedback. *Management Science*, 50(10), 1366–1378.
- Denrell, J., & March, J. G. (2001). Adaptation as information restriction: The hot stove effect. *Organization Science*, 12(5), 523–538.
- Dickson, P. R., & Sawyer, A. G. (1990). The price of knowledge and search of supermarket shoppers. *Journal of Marketing*, 54, 42–53.
- DiMaggio, P. J. (1997). Culture and cognition. *Annual Review of Sociology*, 23, 263–287.
- Douglas, M. (1986). *How institutions think*. Syracuse, NY: Syracuse University Press.
- Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg & J. Davidson (Eds.), *Mechanisms of insight* (pp. 365–395). Cambridge, MA: MIT Press.
- Edmans, A., Garcia, D., & Norli, O. (2007). Sports sentiment and stock returns. *Journal of Finance*, 62, 1967–1998.
- Eisenstein, E. M., & Hutchinson, J. W. (2006). Action-based learning: Goals and attention in the acquisition of market knowledge. *Journal of Marketing Research*, 43, 244–258.
- Faraj, S., & Sprout, L. (2000). Coordinating expertise in software development teams. *Management Science*, 46(12), 1554–1568.
- Gentner, D., & Kurtz, K. (2005). Relational categories. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. W. Wolff (Eds.), *Categorization inside and outside the lab*. (pp. 151–175). Washington, DC: APA.
- Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science*, 10, 277–300.
- George, J. M. (2007). Creativity in organizations. *Academy of Management Annals*, 1, 438–477.
- Gibson, F. P. (2000). Feedback delays: How can decision makers learn not to buy a new car every time the garage is empty? *Organizational Behavior and Human Decision Processes*, 83(1), 141–166.
- Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, 12, 306–355.
- Glynn, M. A. (1994). Effects of work task cues and play task cues on information processing, judgment and motivation. *Journal of Applied Psychology*, 79(1), 34–45.
- Goldenberg, J., Mazursky, D., & Solomon, S. (1999). The fundamental templates of quality ads. *Marketing Science*, 18(3), 333–351.
- Goldstone, R. L., Roberts, M. E., & Gureckis, T. M. (2008). Emergent processes in group behavior. *Current Directions in Psychological Science*, 17, 10–15.

- Goode, M. R., Dahl, D. W., & Moreau, C. P. (2010). The effect of experiential analogies on consumer perceptions and attitudes. *Journal of Marketing Research*, 47, 274–286.
- Gourville, J. T., & Soman, D. (2005). Overchoice and assortment type: When and why variety backfires. *Marketing Science*, 24(3), 382–395.
- Graham, J. R., Harvey, C. R., & Huang, H. (2009). Investor competence, trading frequency, and home bias. *Management Science*, 55(7), 1094–1106.
- Gregan-Paxton, J., & Moreau, C. P. (2003). How do consumers transfer existing knowledge? A comparison of analogy and categorization effects. *Journal of Consumer Psychology*, 13(4), 422–430.
- Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24, 411–435.
- Hargadon, A. B., & Sutton, R. I. (1997). Technology brokering and innovation in a product development firm. *Administrative Science Quarterly*, 42, 716–749.
- Haugvedt, C., Kerr, P., & Kardes, F. (2008). *Handbook of consumer psychology*. Mahwah, NJ: Erlbaum.
- Heath, C., Larrick, R. P., & Klayman, J. (1998). Cognitive repairs: How organizational practices can compensate for individual shortcomings. *Research in Organizational Behavior*, 20, 1–37.
- Heath, C., & Tversky, A. (1991). Preferences and beliefs: Ambiguity and the competence in choice under uncertainty. *Journal of Risk and Uncertainty*, 4, 5–28.
- Higgins, E. T. (2000). Making a good decision: Value from fit. *American Psychologist*, 55, 1217–1230.
- Hirschleifer, D., & Shumway, T. (2003). Good day sunshine: Stock returns and the weather. *Journal of Finance*, 58(3), 1009–1032.
- Hoch, S. J., Bradlow, E. T., & Wansink, B. (1999). The variety of an assortment. *Marketing Science*, 18(4), 527–546.
- Hopkins, P. E. (1996). The effect of financial statement classification of hybrid financial instruments on financial analysts' stock price judgments. *Journal of Accounting Research*, 34, 33–50.
- Jackson, S. E., & Dutton, J. E. (1988). Discerning threats and opportunities. *Administrative Science Quarterly*, 33, 370–387.
- Jeffrey, S. A., Onay, S., & Larrick, R. P. (2010). Goal attainment as a resource: The cushion effect in risky choice above a goal. *Journal of Behavioral Decision Making*, 23, 191–202.
- John, D. R., Loken, B., & Joiner, C. (1998). The negative impact of extensions: Can flagship products be diluted? *Journal of Marketing*, 62, 19–32.
- Juslin, P., Olsson, H., Olsson, A. C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General*, 132(1), 133–156.
- Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.
- Keil, F. C., Stein, C., Webb, L., Billings, V. D., & Rozenblit, L. (2008). Discerning the division of cognitive labor: An emerging understanding of how knowledge is clustered in other minds. *Cognitive Science*, 32, 259–300.
- Keller, J., & Loewenstein, J. (2011). The cultural category of cooperation: A cultural consensus model analysis for China and the US. *Organization Science*, 22, 299–319.
- Keller, J., Loewenstein, J., & Yan, J. (2010). Culturally-guided beliefs about opposing categories and their consequences for action: The case of cooperation and competition. In *Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society* (pp. xx–xx).
- Keller, K. L., & Lehmann, D. R. (2006). Brands and branding: Research findings and future priorities. *Marketing Science*, 25(6), 740–759.
- Kida, T., & Smith, J. F. (1995). The encoding and retrieval of numerical data for decision making in accounting contexts: Model development. *Accounting, Organizations and Society*, 20(7/8), 585–610.
- Kida, T., Smith, J. F., & Maletta, M. (1998). The effects of encoded memory traces for numerical data on accounting decision making. *Accounting, Organizations and Society*, 23(5/6), 451–466.
- Koonce, L., & Mercer, M. (2005). Using psychological theories in archival financial accounting research. *Journal of Accounting Literature*, 24, 175–214.
- Lewis, K., Lange, D., & Gillis, L. (2005). Transactive memory systems, learning, and learning transfer. *Organization Science*, 16(6), 581–598.
- Libby, R., Bloomfield, R., & Nelson, M. W. (2002). Experimental research in financial accounting. *Accounting, Organizations and Society*, 27(8), 775–810.
- Libby, R., Tan, H-T., & Hunton, J. E. (2006). Does the form of management's earning guidance affect analysts earnings forecasts? *Accounting Review*, 81(1), 207–225.
- Loewenstein, J., & Heath, C. (2009). The repetition-break plot structure: A cognitive influence on selection in the marketplace of ideas. *Cognitive Science*, 33, 1–19.
- Loewenstein, J., Raghunathan, R., & Heath, C. (2011). The repetition-break plot structure makes effective television advertisements. *Journal of Marketing*, 75(5), 105–119.
- Loewenstein, J., & Thompson, L. (2006). Learning to negotiate: Novice and experienced negotiators. In L. Thompson (Ed.), *Negotiation theory and research* (pp. 77–98). New York: Psychology Press.
- Loewenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin and Review*, 6(4), 586–597.
- Loken, B. (2006). Consumer psychology: Categorization, inferences, affect, and persuasion. *Annual Review of Psychology*, 57, 453–485.
- Loken, B., & Ward, J. (1990). Alternative approaches to understanding the determinants of typicality. *Journal of Consumer research*, 17, 111–126.
- Lord, R. G., Foti, R. J., & De Vader, C. L. (1984). A test of leadership categorization theory: Internal structure, information processing, and leadership perceptions. *Organizational Behavior and Human Performance*, 34, 343–378.
- MacCormack, A., Rusnak, & Baldwin, C. Y. (2006). Exploring the structure of complex software designs: An empirical study of open source and proprietary code. *Management Science*, 52(7), 1015–1030.
- March, J. G. (1994). *A primer on decision making*. New York: The Free Press.
- Markman, A.B. & Gentner, D. (1993). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language*, 32(4), 517–535.
- Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, 129(4), 592–615.
- Markman, A.B., & Stilwell, C.H. (2001). Role-governed categories. *Journal of Experimental and Theoretical Artificial Intelligence*, 13(4), 329–358.

- McAlister, L., & Pessemier, E. (1982). Variety-seeking trait: An interdisciplinary review. *Journal of Consumer Research*, 9, 311–322.
- McQuarrie, E. F., & Mick, D. G. (1996). Figures of rhetoric in advertising language. *Journal of Consumer Research*, 22, 424–438.
- Meyvis, T., & Janiszewski, C. (2002). Consumers' beliefs about product benefits: The effect of obviously irrelevant product information. *Journal of Consumer Research*, 28, 618–635.
- Meyvis, T., & Janiszewski, C. (2004). When are broader brands stronger brands? An accessibility perspective on the success of brand extensions. *Journal of Consumer Research*, 31, 346–357.
- Moore, D. A., & Flynn, F. J. (2008). The case for behavioral decision research in organizational behavior. *Academy of Management Annals*, 2, 399–431.
- Moreau, C. P., Markman, A. B., Lehmann, D. R. (2001). "What is it?" Categorization flexibility and consumers' responses to really new products. *Journal of Consumer Research*, 27, 489–498.
- Morris, M. W., & Moore, P. C. (2000). The lessons we (don't) learn: Counterfactual thinking and organizational accountability after a close call. *Administrative Science Quarterly*, 45, 737–765.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Engelwood Cliffs, NJ: Prentice-Hall.
- Ocasio, W. (1997). Towards an attention-based view of the firm. *Strategic Management Journal*, 18, 187–206.
- Odean, T. (1998). Volume, volatility, price, and profit when all investors are above average. *Journal of Finance*, 53, 1887–1934.
- Odean, T. (1999). Do investors trade too much? *American Economic Review*, 89, 1279–1298.
- Parsons, J., & Wand, Y. (2008). Using cognitive principles to guide classification in information systems modeling. *MIS Quarterly*, 32(4), 839–868.
- Peng, K., Spencer-Rodgers, J., & Nian, Z. (2006). Naïve dialecticism and the Tao of Chinese thought. In U. Kim, K. S. Yang, & K. K. Huang, (Eds.), *Indigenous and cultural psychology: Understanding people in context* (pp. 247–262). New York: Springer.
- Pennington, N., & Hastie, R. (1986). Explanation-based decision making: Effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(3), 521–533.
- Perry-Smith, J. E. (2006). Social yet creative: The role of social relationships in facilitating individual creativity. *Academy of Management Journal*, 49(1), 85–101.
- Porac, J. F., & Thomas, H. (1994). Cognitive categorization and subjective rivalry among retailers in a small city. *Journal of Applied Psychology*, 79(1), 54–66.
- Porac, J. F., Thomas, H., & Badenfuller, C. (1989). Competitive groups as cognitive communities: The case of Scottish knitwear manufacturers. *Journal of Management Studies*, 26(4), 397–416.
- Porac, J. F., Thomas, H., Wilson, F., Paton, D., & Kanfer, A. (1995). Rivalry and the industry model of Scottish knitwear producers. *Administrative Science Quarterly*, 40(2), 203–227.
- Porac, J. F., Wade, J. W., & Pollock, T. G. (1999). Industry categories and the politics of the comparable firm in CEO compensation. *Administrative Science Quarterly*, 44, 112–144.
- Rahmandad, H., Repenning, N., & Sterman, J. (2009). Effects of feedback delay on learning. *System Dynamics Review*, 25(4), 309–338.
- Ratneshwar, S., Barsalou, L. W., Pechmann, C., & Moore, M. (2001). Goal derived categories: The role of personal and situational goals in category representation. *Journal of Consumer Psychology*, 10, 147–157.
- Read, D., Loewenstein, G., & Rabin, M. (1999). Choice bracketing. *Journal of Risk and Uncertainty*, 19(1–3), 171–197.
- Rindova, V. P., & Petkova, A. P. (2007). When is a new thing a good thing? Technological change, product form design, and perceptions of value for product innovations. *Organizational Science*, 18(2), 217–232.
- Rosa, J. A., Judson, K. M., & Porac, J. F. (2005). On the sociocognitive dynamics between categories and product models in mature markets. *Journal of Business Research*, 58, 62–69.
- Rosa, J. A., Porac, J. F., Runser-Spanjol, & Saxon, M. S. (1999). Sociocognitive dynamics in a product market. *Journal of Marketing*, 63, 64–77.
- Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(4), 629–639.
- Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, 70, 27–52.
- Saunders, E. M. J. (1993). Stock prices and wall street weather. *American Economic Review*, 83, 1337–1345.
- Schweitzer, M. E., Ordoñez, L. D., & Duouma, B. (2004). Goal setting as a motivator of unethical behavior. *Academy of Management Journal*, 47(3), 422–432.
- Sedor, L. M. (2002). An explanation for unintentional optimism in analysts earnings forecasts. *The Accounting Review*, 77(4), 731–753.
- Seybert, N., & Bloomfield, R. (2009). Contagion of wishful thinking in markets. *Management Science* 55(5), 738–751.
- Simon, H. A. (1947). *Administrative behavior*. New York: The Free Press.
- Simon, H. A. (1955). A behavioral model of rational choice. *Journal of Economics*, 59, 99–118.
- Simon, H. A. (1969). *Sciences of the artificial*. Cambridge, MA: MIT Press.
- Simonson, I. (1990). The effect of purchase quantity and timing on variety-seeking behavior. *Journal of Marketing Research*, 27(2), 150–162.
- Singley, M., & Anderson, J. (1989). *The transfer of cognitive skill*. Cambridge, MA: Harvard University Press.
- Sorescu, S., & Subrahmanyam, A. (2006). The cross section of analyst recommendations. *Journal of Financial and Quantitative Analysis*, 41(1), 139–168.
- Sprinkle, G. B., & Williamson, M. G. (2006). Experimental research in managerial accounting. In C. S. Chapman, A. G. Hopwood, & M. D. Shields (Eds.), *Handbook of management accounting research* (Vol. 1, pp. 415–444). New York: Elsevier.
- Sterman, J. D. (1989a). Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management Science*, 35, 321–339.
- Sterman, J. D. (1989b). Misperceptions of feedback in dynamic decision making. *Organizational Behavior and Human Decision Processes*, 43, 301–335.
- Subrahmanyam, A. (2007). Behavioural finance: A review and synthesis. *European Financial Management*, 14(1), 12–29.

- Taylor, A., & Greve, H. R. (2006). Superman or the Fantastic Four? Knowledge combination and experience in innovative teams. *Academy of Management Journal*, 49(4), 723–740.
- Taylor, S., & Todd, P. (1995). Assessing IT usage: The role of prior experience. *MIS Quarterly*, 19(4), 561–570.
- Te'eni, D. (2001). A cognitive-affective model of organizational communication for designing IT. *MIS Quarterly*, 25(2), 251–312.
- Tenbrunsel, A. E., & Messick, D. M. (1999). Sanctioning systems, decision frames, and cooperation. *Administrative Science Quarterly*, 44, 684–707.
- Tenbrunsel, A. E., & Smith-Crowe, K. (2008). Ethical decision making: Where we've been and where we're going. *Academy of Management Annals*, 2, 545–607.
- Thagard, P., & Shelley, C. P. (2001). Emotional analogies and analogical inference. In D. Gentner, K. H. Holyoak, & B. K. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 335–362). Cambridge, MA: MIT Press.
- Thibodeau, J. C. (2003). The development and transferability of task knowledge. *Auditing: A Journal of Practice and Theory*, 22(1), 47–67.
- Todd, P., & Benbasat, I. (1992). The use of information in decision-making: An experimental investigation of the impact of computer-based decision aids. *MIS Quarterly*, 16(3), 373–393.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131.
- Vakratsas, D., & Ambler, T. (1999). How advertising works: What do we really know? *Journal of Marketing*, 63(1), 26–43.
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425–478.
- Walsh, J. P., & Ungson, G. R. (1991). Organizational memory. *Academy of Management Review*, 16(1), 57–91.
- Wang, W., & Benbasat, I. (2009). Interactive decision aids for consumer decision making in e-commerce: The influence of perceived strategy restrictiveness. *MIS Quarterly*, 33(2), 293–320.
- Ward, T. B. (1994). Structured imagination: The role of category structure in exemplar generation. *Cognitive Psychology*, 27(1), 1–40.
- Ward, T. B., Patterson, M. J., & Sifonis, C. M. (2004). The role of specificity and abstraction in creative idea generation. *Creativity Research Journal*, 16(1), 1–9.
- Watson-Manheim, M. B., & Bélanger, F. (2007). Communication media repertoires: Dealing with the multiplicity of media choices. *MIS Quarterly*, 31(2), 267–293.
- Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives*, 18(2), 107–126.
- Wood, R. E. (1986). Task complexity: Definition of the construct. *Organizational Behavior and Human Decision Processes*, 37, 60–82.
- Yeung, C. W. M., & Soman, D. (2007). The duration heuristic. *Journal of Consumer Research*, 34, 315–326.

William Forde Thompson and Paolo Ammirante

Abstract

Listening to music entails processes in which auditory input is automatically analyzed and classified, and conscious processes in which listeners interpret and evaluate the music. Performing music involves engaging in rehearsed movements that reflect procedural (embodied) knowledge of music, along with conscious efforts to guide and refine these movements through online monitoring of the sounded output. Composing music balances the use of intuition that reflects implicit knowledge of music with conscious and deliberate efforts to invent musical textures and devices that are innovative and consistent with individual aesthetic goals. Listeners and musicians also interact with one another in ways that blur the boundary between them: Listeners tap or clap in time with music, monitor the facial expressions and gestures of performers, and empathize emotionally with musicians; musicians, in turn, attend to their audience and perform differently depending on the perceived energy and attitude of their listeners. Musicians and listeners are roped together through shared cognitive, emotional, and motor experiences, exhibiting remarkable synchrony in behavior and thought. In this chapter, we describe the forms of musical thought for musicians and listeners, and we discuss the implications of implicit and explicit thought processes for musical understanding and emotional experience.

Key Words: music, performance, composition, improvisation, emotion, listening, mood, arousal, rhythm, expression, timing, melody, harmony

Introduction

Music is a puzzling phenomenon because its manifestations, uses, and social functions are highly variable. Like language, music unfolds in time and varies in pitch, timing, timbre, and intensity, and music performance, as with speaking, is accompanied by meaningful gestures and facial expressions that contribute to its intelligibility. Unlike language, musical behavior often seems discretionary. If language were to disappear, there would be profound and catastrophic consequences for virtually every facet of our lives. The disappearance of music would be an unwelcome change, but the world could continue roughly in the same manner as before.

Perhaps because music has no one function that is essential to human survival, it is free to take many

different forms within and across cultural groups. Among its varied uses are aesthetic enjoyment, emotional communication, symbolic representation, social conformity, and group bonding. It is also often used to accompany or encourage synchronous group movement (e.g., dancing, marching) or as symbolic validation of social institutions and religious rituals. Many musicologists are reluctant to assume that there are essential structural or functional properties of all music, and music is often defined broadly. For example, the *American Heritage Dictionary* defines music as the “arranging of sounds in time.” Such a definition is highly inclusive but seems unsatisfactory in ignoring the aesthetic, emotional, and motional properties of music. Ian Cross (2003a, 2003b) carefully considered the problems

of defining music and argued that it always involves the creation of sounds and movements that have the potential to carry meaning across a range of contexts. This latter characteristic, what Cross calls *floating intentionality*, allows music to be exploited for a range of purposes.

If music has no essential function, then why is musical behavior observed in every known culture, and why is it so important to us? In Western industrialized cultures, people spontaneously sing while engaged in activities such as house cleaning, driving, or showering. Live or recorded music is played at social gatherings such as funerals, graduation ceremonies, birthday parties, and weddings. Young adults often form passionate attachments with specific musical styles and may nurture a sense of identity that is deeply connected with specific groups of musicians (whether followers of The Grateful Dead or Lady Gaga). The cognitive and educational costs involved in music appreciation and production are high. Music is taught at home and all levels of education, and society places enormous value on the craft of music making—the performers and composers of music. Acquiring expertise as a performer or composer can take 10 or more years of intensive training (roughly 10,000 hours of effortful practice) and elite musicians are among the most esteemed members of a society.

In this chapter, we review current knowledge of the psychological basis of music, focusing on cognitive processes involved in the activities of music listening, performing, and composing. These processes are best described as music relevant rather than music specific. That is, they handle input from any domain, but a breakdown in that process is more likely to affect music perception than the perception of speech or environmental sounds because music entails fine-grained changes in acoustic attributes (Peretz & Hyde, 2003). Indeed, the severe musical deficits observed in individuals with tone deafness or “amusia” may arise from processes that are not specialized for musical input (Douglas and Bilkey, 2007; Nan, Sun, & Peretz, 2010; Patel, Wong, Foxton, Lochy, & Peretz, 2008). We also explore the relevance of cognitive constraints, involuntary processes, and focused attention for musical activities, addressing the following questions: To what extent does music behavior rely on conscious, explicit processes? To what extent does it rely on early predispositions, unconscious biases, and other involuntary processes? Can people acquire knowledge of musical structure in the absence of formal training? Why

does music give rise to such powerful aesthetic and emotional experiences? To what extent do perceptual and cognitive processes constrain our capacity to perceive, perform, and compose music?

A central theme of the chapter is that listeners, performers, and composers are “roped together” by shared cognitive and motor processes. The distinctions made between these three activities are not always clear cut. Listeners are most often active participants in music making (singing along, tapping to the beat). Performers and composers are almost invariably listeners (a notable exception is the deaf percussionist Evelyn Glennie). As such, the sections on performing and composing in this chapter make frequent reference to music listening, illustrating, for example, how the mechanism of auditory scene analysis called auditory streaming can shape performance and compositional choices. Other well-studied mechanisms that may have served the same illustrative purpose, such as listener sensitivity to consonance and dissonance between tones, are exhaustively reviewed elsewhere (Bregman, 1990; Handel, 1989). Approaches to cognition in music performance and composition that are less directly related to the listener are also covered elsewhere and include musical expertise (Lehmann & Gruber, 2006; Sloboda, 2000; Sloboda, Davidson, Howe, & Moore, 1996), improvisation (Pressing, 1984, 1988) and composition (McAdams, 2004; Sloboda, 2004) synchronization in ensemble playing (Goebl & Palmer, 2009; Rasch, 1988), and sensorimotor integration in music performance (Goebl & Palmer, 2008; Palmer, Koopmans, Loehr, & Carter, 2009).

Listening to Music

The specific arrangements of sound that we perceive and characterize as music are referred to here as *structures*. In Western music, these include individual sequences of tones perceived as *melodies* (melodic structure), simultaneous combinations of tones perceived as *chords* (harmonic structure), and distinctive temporal patterns of events perceived as *rhythms* (rhythmic structure). Structures can be remarkably complex, especially in the domains of pitch and rhythm. Yet with no formal music training or vocabulary to describe musical structure, listeners readily appreciate and respond to music, often without knowing why. Sequences of pitches may be perceived as a *melody*, but not all pitch sequences are heard as such. Indeed, the term “melody” is a *psychological* construct rather

than a physical description of a sonic event. Melody perception is dependent on perceptual mechanisms that group sequences of tones into a coherent stream and segregate them from sounds that occur before, after, or concurrently with those tones. Auditory stream segregation is important for music perception because melodies are often accompanied by other textures in the music, forming simultaneous combinations of pitches that are perceived as the *harmony* of the music. Not all tone combinations are perceived as harmonious and pleasing: Like melody, harmony is a psychological construct that depends on a perceptual interpretation of sound waves impinging on the eardrum.

Melody and harmony unfold rhythmically and often suggest a regular cycle of strong and weak accents or “beats,” collectively called the meter. Rhythm and meter, like melody and harmony, are psychological constructs because their appreciation depends on the operation of perceptual and cognitive processes. Because of these processes, one can perceive a strong metric accent at a time in the music when there is no sound. Once a metric framework is established, the brain “infers” subsequent cycles of strong and weak accents through expectancy mechanisms. Important musical events tend to coincide with metrically strong beats, so listeners anticipate points of metric stress and allocate greater attentional resources to them. As such, our attention to music occurs not in an evenly distributed manner over time but exhibits cyclical characteristics described as *dynamic attending* (Jones, 1989; Large & Jones, 1999). The cyclical nature of attention means that some musical elements (coinciding with strong accents) are more salient and better remembered than others (coinciding with weak accents). Dynamic attending need not be limited to meter. For example, by spanning a regular number of metrical cycles, melody (e.g., four-bar phrasing) and harmonic progression (e.g., twelve-bar blues) may generate further expectancies at longer cycles. Multiple cyclical boundaries give music a nested or hierarchical quality (Lerdahl & Jackendoff, 1983). Other attributes of sound also vary systematically in music, such as event density, intensity, timbre, and pace or tempo. These attributes also influence the dynamic nature of auditory attention and the perceived hierarchical structure of music. Musicians manipulate almost all perceivable attributes of sound, and all perceivable attributes contribute to a listener’s perception and appreciation of music.

Perceiving Musical Structure

For music to acquire meaning, listeners must perceive its structure and organization. Musical meaning can arise from relationships between adjacent and nonadjacent events (formalist meaning) and from the feelings or emotions that these relationships evoke (expressionist meaning; Meyer, 1956). A first step in this process is to differentiate musical sound from meaningless noise and other nonmusical sounds. Performers and composers—the producers of music—intuitively take steps to facilitate the process of perceiving musical structure. Performers introduce expressive actions that highlight or enhance structural features, slowing down at the end of phrases, accenting musically significant notes, and introducing timing asynchronies so that melodic lines stand out from the accompanying musical material. Most composers create music with structures that can be easily perceived and remembered. They often introduce a metric framework so that listeners can keep track of strong and weak accents, and they manipulate the musical materials so that some events are strongly associated with closure and stability while others connote instability and tension. Once established, these associations provide listeners with a means of understanding the musical narrative: the increases and decreases in tension, momentary points of repose, unexpected outcomes, and a sense that the music has come to an end (Cuddy & Lunney, 1995; Lerdahl & Jackendoff, 1983; Meyer, 1956; Narmour, 1992; Schellenberg, 1997). Of course, some composers intentionally avoid such conventions (e.g., Schoenberg, Webern), but such music does not enjoy widespread appreciation.

The attributes and structures that comprise music are influenced by the architecture of human cognition, providing a source of commonality in the forms of music that are most readily perceived and remembered across cultures, historical periods, and genres. For example, the limits of working memory exert an impact on optimal phrase lengths in melodies across cultures. This constraint does not mean that long phrases are impossible to appreciate, but most people across cultures prefer melodic phrases that can be readily perceived and remembered as a coherent musical package. Music can be appreciated on multiple levels of analysis, including local structure (tones, intervals), higher order structure (phrases, movements), and abstract structure (scales, keys).

Our ability to perceive musical structure is partly driven by general mechanisms such as pitch perception and auditory scene analysis, and partly shaped

by long-term knowledge of abiding regularities in music. This knowledge can be gained through formal training in music, but it arises involuntarily following long-term exposure to music by a process of implicit learning. It is through implicit learning that most Western listeners acquire their knowledge of Western tonal music (such as the music of Mozart and Bach, but including contemporary popular and folk genres). The same processes allow non-Western listeners to acquire implicit knowledge of the music of their culture, whether the Carnatic music of South India, the Sami music of Northern Finland, or Indonesian Gamelan music.

Following passive exposure to music, listeners acquire extensive knowledge of the music of their culture, allowing them to anticipate, evaluate, and respond to its characteristic features. In most cultures, music is highly structured and requires complex mental representations for its appreciation. Western tonal music is a system based on a relatively small number of elements that exhibit high regularity in their occurrence and co-occurrence. It is based on a set of 12 pitch classes, from which a subset of seven notes typically dominates the music (for example, the seven notes of the major scale: doh-re-mi-fa-sol-la-ti). Sequences of notes from this subset form the melodies of music. In turn, chords are defined by sets of notes played simultaneously. For example, the notes of C, E, and G played simultaneously define a C major chord. Sequences of chords are the basis of musical harmony. Through persistent exposure to the music in one's environment, listeners internalize regularities in the melodic and harmonic materials to which they are exposed. These regularities include typical levels of musical stability or tension associated with different scale notes and chords, and customary transitions between the notes and chords that make up melodic and harmonic sequences.

Acquiring Sensitivity to Musical Structure

Because musical regularities vary across cultures, infants and young children must abstract the specific properties of the systems in their environments. This process takes many years. Its development can be traced in preverbal infants as an emerging sensitivity to subtle changes in a repeated melody, as reflected in measures such as head-turn responses, looking times, and differential sucking rates. Stimuli that are perceived to be novel elicit changes in these measures. Thus, by varying the degree of novelty or familiarity of a stimulus, and then observing

changes in these behaviors, researchers can evaluate sensitivity to various aspects of musical structure. Before 1 year of age infants show little sensitivity to the features of Western tonal conventions, having insufficient time to abstract the relevant features. For example, they detect changes to a melody equally regardless of whether the new note conforms to the implied harmony or scale of the original melody (Schellenberg & Trehub, 1999; Trainor & Trehub, 1992). By the age of 4 years, they more easily detect changes to a diatonic melody (i.e., made up of conventional scale notes) than changes to a nondiatonic melody (Trehub, Cohen, Thorpe, & Morrongiello, 1986). By the age of 7 years, children exhibit sensitivity to Western harmony (Trainor & Trehub, 1994). How do infants and children acquire this knowledge of Western tonal conventions?

Saffran, Johnson, Aslin, and Newport (1999) provided evidence that listeners acquire sensitivity to statistical properties of music through passive listening. They constructed a small artificial musical "vocabulary" consisting of six, three-note sequences or "tone-words." They then constructed a 7-minute sequence by chaining together these tone-words and repeating them at varying temporal locations. The result was a continuous stream of tones in which certain tone transitions occurred more often than others. Tone transitions that occurred in the tone-words occurred repeatedly and were therefore associated with high transitional probabilities (average $p = .64$), just as the probabilities of phoneme transitions that occur in linguistic words are high. In turn, tone transitions that did not occur in any of the tone-words had low transitional probabilities (they could occur between the end of one figure and the start of another, average $p = 0.14$).

Twenty-four nonmusicians heard the 7-minute sequence three times, unaware that the sequence was constructed from a vocabulary of three-note tone-words. After the exposure phase they were presented pairs of three-note stimuli in the test phase. For each pair, one stimulus was a tone-word and the other was a novel (nonword) tone sequence. Listeners indicated which of the two figures was most familiar. As expected, figures that were part of the vocabulary of the 7-minute sequence were judged as significantly more familiar to listeners than were the novel figures. All six tone-words were learned at a level significantly better than chance. Although none of the tone-words were heard in isolation prior to the test phase, listeners internalized the high predictability of note transitions in the tone-words and

perceived them as familiar. Further analyses suggested that the impression of familiarity was based on sensitivity to the conditional probabilities of tone transitions in the tone-words.

MUSICAL PREFERENCES

The aforementioned results indicate that sensitivity to transitional probabilities can lead to an impression of familiarity. Research on the “mere exposure effect” illustrates that statistical learning also affects our preferences for music. Wilson (1979, Experiment 2) repeatedly exposed participants to melodic sequences in one ear while presenting spoken material to the other ear (exposure phase). Participants were required to shadow the verbal material (say each word aloud as they heard it) while proofreading written text of the verbal material. In a subsequent (test) phase, they rated the exposed and novel melodic sequences for preference and indicated whether they remembered hearing each sequence in the proofreading phase. Recognition of the exposed sequences was not above chance but preference was significantly higher for exposed than for novel sequences.

Peretz, Gaudreau, and Bonnel (1998) provided further evidence for this effect. Listeners were exposed to a set of familiar and unfamiliar melodies (exposure phase). They were then presented with exposed and novel melodies (test phase). Half of the listeners rated their preference for the test melodies (implicit test) and half identified the melodies that they remembered hearing in the exposure phase (explicit test). Both preference for and recognition of exposed melodies were above chance, but when the investigators varied the time interval between exposure and test phases (between-task interval), they observed a different temporal gradient of memory loss for preference and recognition judgments, suggesting that recognition and preference arise from independent memory systems. Changes in timbre between exposure and test phases (piano vs. flute tones) did not affect preference judgments but negatively impacted recognition, which further implies independence between recognition and preference, and illustrates the robustness of exposure effects on preference.

In short, statistical learning leads to impressions of familiarity and preference, effects that are arguably forms of implicit memory (see Evans, Chapter 8; Zajonc, 1968). Thompson, Balkwill, and Vernescu (2000) found evidence that recently heard melodic patterns also shape the expectations that listeners

form about notes to follow, beyond what can be explained by explicit memory for those patterns. Listeners were presented with 30 short tone sequences (target sequences) three times each (exposure phase) and reported the number of notes in each. They were then presented target and novel sequences and rated how well the final note continued the pattern of notes that preceded it. Novel sequences were identical to target sequences except for the final note. Ratings were significantly higher for target sequences than for novel sequences, indicating that mere exposure to melodic sequences influences expectations about future events in music. These effects could not be explained by explicit memory for the target sequences or by anything inherently predictable about the target sequences.

Emotional Responses to Music

Research on statistical learning illustrates how general cognitive processes can shape music experiences, giving rise to feelings of familiarity, preference, and expectation as music unfolds. The effects of passive listening on expectation are particularly significant because violations and fulfillments of expectancies are responsible for some of the most powerful effects of music. Meyer (1956) was among the first to discuss the role of expectancies in emotional responses to music. When we listen to music that is culturally familiar, we continuously form expectations about what will occur, and when (Huron, 2006; Thompson, Cuddy, & Plaus, 1997). Composers and performers toy with our expectations, fulfilling our expectations in some cases but also delaying or thwarting expected events at other moments. Composers manipulate expectancies primarily through the attributes of pitch and temporal structure (Thompson & Stainton, 1998). Performers use expressive devices to manipulate expectancies, such as crescendo (a gradual increase in loudness), tempo (the overall pace of the music), and ritardando (a gradual slowing down).

Violations to expectations are charged emotionally because the mechanisms for predicting events are central to all behavior and indeed to human survival. To function, people must continuously anticipate the future, whether in the simple act of walking (knowing when your feet will contact the ground), planning a meal, avoiding danger, or interacting socially. The feeling of surprise at an unexpected outcome is a biological signal that current strategies of prediction have gone awry and need refinement (Mandler, 1984).

Responses to violations and fulfillments of expectations in music can account for a wide range of emotional experiences, not just surprise. Huron (2006) outlined some of the ways that expectancies generate complex emotional responses. He pointed out that expectations occur on different levels of analysis, from physiological preparation for an imminent event, to longer term expectations of future states. He also posited different types of reactions to expected outcomes. The *prediction response* is a transient state of reward or punishment that arises in response to the accuracy of prediction. Accurate expectations are encouraged by positively valenced emotional responses; inaccurate expectations are discouraged by negatively valenced emotional responses. The *reaction response* is a rapid, automatic bodily or visceral reaction based on an immediate assessment of an event, independent of whether it has been accurately predicted. The *appraisal response* is a more reflective and conscious evaluation of an outcome, and it need not be compatible with the reaction response.

Expectancies for music, formed primarily by statistical learning, generate powerful emotional responses to music (e.g., happiness, sadness). Other attributes of music, including overall intensity, pitch height, and tempo, also have emotional connotations independently of expectancy mechanisms, possibly because of the association between these attributes and emotion in human speech (Ilie & Thompson, 2006; Juslin & Laukka, 2003). Ilie and Thompson (2011) presented eight groups of listeners with 7 minutes of music by Mozart. All groups received the same composition, but the music was manipulated in tempo (slow, fast), pitch height (low, high) and intensity (soft, loud). Manipulations influenced three dimensions of affective experience: energetic arousal, tension arousal, and valence. For example, tension arousal was higher after listening to loud music than soft music; energetic arousal was higher after listening to fast music than slow music; and valence was higher after listening to high-pitched music than low-pitched music. When these manipulations were applied to a narrated documentary, they gave rise to the same affective consequences, illustrating that music and speech may share a common emotional code.

The same manipulations also affected cognitive function. Exposure to fast music not only led to the experience of high energetic arousal; it enhanced speed of processing (as measured by reaction time in a routine task). Exposure to high-pitched music

not only led to a positively valenced experience, it also enhanced creative problem solving, as measured by success on Dunker's (1945) candle problem and Maier's (1931) two-string problem (see van Steenburgh et al., Chapter 24). In short, listening to music has multiple effects on thought and experience. Listeners internalize statistical properties of the music and develop preferences, feeling of familiarity, and expectancies for common tone transitions. They also perceive and experience strong emotions, and their cognitive functioning may be significantly affected.

Performing Music

With decades of intensive training, practicing musicians perform highly complex passages of music with apparent ease. Elite singers, for example, exercise precise vocal control so as to manipulate sounded attributes such as pitch, intensity, vibrato, timing, and timbre. To what extent do music performers rely on implicit and explicit processes in order to execute a performance? Performers effortlessly produce highly practiced movement sequences that reflect embodied or procedural knowledge of music. At the same time, they monitor the sounded output of their actions, consciously reflecting on the music that they are producing and making online adjustments to optimize their communicative intentions.

To illustrate the delicacy of this balance, imagine an established singer performing an aria from Mozart's *The Marriage of Figaro*. She has performed the aria innumerable times and typically receives standing ovations for her performances. In one performance at the prestigious *Lyric Opera* in Chicago, she is approaching a particularly complex and beautiful vocal passage in which the sung melody mirrors the accompanying violins. Unexpectedly, her online reflections on the performance begin to intrude excessively into her consciousness, disrupting the delicate balance between implicit and explicit processes required for an optimal performance. Although potentially disastrous, the singer cannot help but worry about the difficulty of an upcoming passage, and she entertains the possibility of a lapse in memory. Suddenly, a performance that has been delivered effortlessly in hundreds of previous performances ceases to be fluent and automatic, and the singer is filled with self-doubt and anxiety. To flounder in front of thousands would be catastrophic, and her conscious awareness of this possibility makes matters worse. From that moment

onward, each sung note feels effortful and tenuous, and there is a very real possibility that her performance will break down.

The musician in question was Renée Fleming, arguably one of the most successful opera singers today (Fleming, 2005; see also Lehrer, 2009, p. 132). Such instances of music performance anxiety are remarkably common in both experienced and inexperienced musicians (Kenny, 2010). The widespread prevalence of music performance anxiety illustrates that all musicians are vulnerable to disruption of the balance between the use of procedural memory and conscious processes that musicians must rely on to guide a performance.

Why is procedural memory important for performers? One reason is that working memory has insufficient capacity to process the amount of information involved in a musical performance. Performing long pieces of music places enormous demands on memory (Aiello & Williamon, 2002) and in the absence of external cues (sheet music) this load is best handled in procedural memory. Procedural memory can be used to generate rapid pitch (or chord) sequences that are implemented with precise timing and expressive actions such as the *ritardando* (a gradual slowing down), *accelerando* (a gradual speeding up), *crescendo* (a gradual increase in loudness), and *diminuendo* (a gradual decrease in loudness). That is, procedural memory not only can be used to store the motor commands needed to perform a sequence of carefully timed notes; it can also be used to store the various changes in timing, pitch, and loudness that are collectively known as *performance expression*.

Understanding Performance Expression

Performance expression refers to those qualities of an individual performance that distinguish it from a strict or mechanical rendering of the set of pitches and durations that are (or could be) indicated in a written or “notated” representation of that music (called a score), and that permit different expressive interpretations of the same piece of music. In Western classical genres, a score provides a reduced or idealized representation of the music. It outlines the basic sequence of notes to be played (pitches and durations), a structural framework in the form of a key signature and a time signature, and selected guidelines for the performer at specific points in the music. Such expressive markings include desired loudness of notes or phrases using abbreviations such as *f* for the Italian *forte*

(loud), *ff* or *fff* to indicate greater intensity levels, *pp* to indicate *pianissimo* (soft), and so forth. Performance expression can be understood as those qualities of the music that go beyond the set of pitches and durations (the notes) as indicated in the score. It includes variations in timing (e.g., rubato, accelerando, ritardando), intensity (e.g., accent, crescendo, diminuendo), timbre, and articulation (e.g., legato, staccato). If music were played without performance expression, it would sound mechanical and regular, and it would likely sound less expressive and emotional.

The expressive actions of performers are deliberate and systematic, and they help listeners to perceive structural features in the music, anticipate climactic moments, grasp emotional connotations, and appreciate points of tension and relaxation. Expressive devices highlight phrase boundaries, harmonic changes, points of accent, and other structural attributes (Ashley, 2002). For example, performers may use intensity and timing to accent notes that have a significant musical function, and they may insert pauses at phrase boundaries to enhance perceptual grouping. Timbre and pitch can also be manipulated in expressive ways. By increasing force to the bow, a violinist can increase the amount of noise at the onset of a note. Similarly, guitar distortion is often used in rock music to suggest friction and instability. When music is performed well, listeners are in a better position to form a stable representation of the structural and emotional properties of that music (Clarke, 1988; Palmer, 1997, 1989; Thompson & Cuddy, 1997).

For elite performers, expressive actions are stored in procedural memory and can be implemented with little or no conscious effort. Conscious efforts to play with no expression do not eliminate them entirely, suggesting that expressive actions become automatized after extensive practice (Bengtsson & Gabrielsson, 1983; Palmer, 1989; Seashore, 1938). Consistent with such results, some researchers surmise that conscious processes are most relevant only for beginners, or when one is learning something new; experts are capable of drawing upon implicit knowledge with no need for conscious thought (Ericsson, 2006). This reliance on implicit processing and skirting of explicit cognition is a hallmark of expertise. As a case in point, when musicians are practicing a new piece they need to make many conscious decisions about technical issues such as fingering and body position, as well as issues of interpretation, such as how to implement expressive actions

such as phrasing, crescendi, rubati, and pedaling at various points in the music. The implementation of such conscious decisions becomes largely automatic with extensive practice, reducing demands on conscious processes during a performance (Lehmann & Gruber, 2006).

Others, however, point out that the flexibility that elite performers exhibit cannot be explained by the exclusive use of procedural memory, and that performers must draw upon a declarative “mental road map” held in long-term working memory (Chaffin & Imreh, 2002; Chaffin, Imreh, & Crawford, 2002; for a discussion of this view, see Geeves, Christensen, Sutton, & McIlwain, 2008). This mental road map is particularly important when something goes wrong in a performance: Under such circumstances a performer “must know where he or she is in the piece, and be prepared to put the performance back on track” (Chaffin & Imreh, 2002, p. 342). Conceptual memory is vital for restarting a motor sequence, and for monitoring the pace, precision, and expressive quality of one’s performance. In turn, performers pay particular attention to their technical and aesthetic goals during difficult or climactic moments in a piece.

The difficulty of separating deliberate and automatic processes is exemplified in the many perceptual interactions that occur between attributes of music. Such interactions are largely outside of the awareness of musicians and listeners but exert a significant influence on music performance, experience, and memory. Conscious attention may be directed toward one attribute, but responses and evaluations to that attribute are often biased by many other attributes. Our experience of a change in pitch (called an *interval* when the change is between successive discrete tones) is influenced by a range of ostensibly irrelevant attributes such as overall pitch height, intensity, and timbre (Russo & Thompson, 2005a; 2005b). The facial expressions of singers also affect how listeners perceive the music (Thompson, Graham, & Russo, 2005) even when they are instructed to focus exclusively on the acoustic information (Quinto, Thompson, Russo, & Trehub, 2010; Thompson, Russo, & Quinto, 2008). Performers, in turn, preattentively implement a range of facial movements in live performances to clarify or enhance emotional messages in the music (Livingstone, Thompson, & Russo, 2009) and to highlight musical structure (Thompson & Russo, 2007; Thompson, Russo, & Livingstone, 2010).

Pitch and Timing

The timing of musical events is biased by the pitches of those events, an interaction that affects performers and listeners (Ammirante, Thompson, & Russo, 2011; Prince, Schmuckler, & Thompson, 2009). One explanation for this bias is that it reflects an implicit awareness of the spatiotemporal correlations in moving bodies in the natural environment (Helmholtz, 1962/1867; Jones & Huang, 1982). To change directions, moving bodies must slow down. When a melody changes direction (e.g., from upward to downward pitch motion), does the music itself seem to slow down? Boltz (1998) found that tempo judgment negatively covaried with the number of changes in pitch direction in isochronous melodies, implying that, just as with moving bodies that change direction, a musical sequence seems to lose momentum at changes in pitch direction.

Ammirante and colleagues predicted that tonal motion implied by pitch structure is mirrored in concurrent motor activity (Ammirante & Thompson, 2010; Ammirante, Thompson, & Russo, 2011). Participants maintained a steady beat by tapping on a single key of a MIDI keyboard. Each tap triggered a discrete tone in a sequence varying in pitch. Although participants were instructed to focus on tapping and ignore the tones that were triggered by their taps, the tonal information significantly affected performance on the tapping task. Where pitch structure implied faster tonal motion (e.g., unidirectional pitch motion, larger pitch intervals “traversed” between tones), the intertap interval (ITI) initiated by the just-triggered tone and the velocity of the tap that followed (TV) were faster; where pitch structure implied slower tonal motion (e.g., a change in pitch direction, smaller pitch intervals), slower ITI and TV followed. These findings suggested that participants failed to disambiguate velocity implied by tonal motion from the velocity of their finger trajectory.

Ammirante et al. argued that the perceived relation between pitch movement and speed (tempo) unconsciously induced local deviations in timing and loudness (a louder sound is produced when a key is struck with faster TV). Interestingly, such temporal and dynamic deviations are also found in expressive music performance, suggesting that certain expressive actions by musical performers reflect sensitivity to spatiotemporal correlations in the natural environment. In effect, music is unconsciously perceived to have “life-like” qualities that give rise to perceptual interactions between musical features. These

interactions are evident in perceptual judgments and timed movements, blurring the distinction between performers and listeners. Its external manifestation may be unconsciously or deliberately communicated in expressive music performance (Truslit in Repp 1993/1938) or “corporeally articulated” in the unconscious actions of an engaged listener, such as swaying with the music (Leman, 2009).

Composing Music

Elite music performance demands flexibility and balance between implicit and explicit processes in real time; composition entails a comparable flexibility and balance applied over a longer time scale. Unlike music performance, where a performer often takes considerable guidance from a musical score, there are few boundaries on the process of music composition. How do composers narrow down the set of decisions involved in such an ill-defined task? One approach is to compose music within the boundaries of a standard musical form such as sonata, 12-bar blues, musical theater, or opera. Composers are also conscious of the biophysical constraints on what can be physically realized by a performer, and the acoustic effects of combining sounds.

For the most part, music is composed with listeners in mind. As such, compositional choices are influenced by an awareness of *musical accessibility*. This includes culturally acceptable ways that the conventions of a stylistic idiom may vary. Music creators may attempt to strike a balance between producing music that is familiar but not trite, and that expands a stylistic idiom without being opaque. One approach is to manipulate expected features. For example, the rhyming style of the “new school” of rap that emerged in the late 1980s replaced simple, monosyllabic end rhymes that coincided with the fourth beat of a four-beat metric cycle with low-probability multisyllabic rhymes and an expressive dodging of the onset of the fourth beat (Bradley, 2010).

Tonal Structure

In the evolution of the Western classical composition, tonality is a distinguishing feature. Western listeners are highly sensitive to tonality, and they expect it. Indeed, most listeners find music that does not have a tonal center confronting or disturbing (e.g., the atonal music of Schoenberg, the free jazz of Ornette Coleman). What accounts for this rigid expectation for tonality? Two mental representations,

both implicit forms of knowledge, may underlie it. The first maps relationships between tones within a musical framework called the *key*. The second maps relationships between keys. Listeners acquire these representations through passive exposure to tonal music (e.g., by listening to music such as Mozart, Bach, Haydn, and most Western popular genres). Once these representations are established, a sense of tonality rapidly emerges while listening to music, as long as the music implies a collection of tones that have a hierarchical organization (Cohen, 1991). The most common tone collection in Western music is the major scale—doh-re-mi-fa-sol-la-ti. Within that scale, the first note (called the tonic) is the most stable pitch and often occurs at points of stability, such as phrase endings or down beats. Other pitches have subordinate levels of stability. Guided by the second representation, listeners are also sensitive to relationships among keys. For example, the key of C is strongly related to the key of G but weakly related to the key of F#. Movement between related keys sounds natural and may even go unnoticed; movement between distantly related keys can sound unnatural or abrupt.

Stylistic changes between the 18th and 20th centuries are characterized by a conscious destabilization of a central recurring tone or “tonic” around which compositions are organized. Composers were troubled by their unconscious adherence to tonal compositions and worried that they would repeatedly “reinvent the wheel” unless they consciously attempted to avoid clear tonal centers. Thus, they began to compose music with an increased emphasis on tones that were distantly related to the tonic, and they shifted the tonic within and between musical movements. These strategies culminated in the works of the late 19-century Romantics such as Wagner and Mahler. In the early 20th century, serialism employed compositional procedures that more rigorously avoided the establishment of any tonal centre or hierarchical organization. One form of serial composition—12-tone music—uses a “tone row” as a musical theme. As a rule, each member of the set of discrete pitch classes contained in the tone row (the 12 tones of the chromatic scale) must be presented before a member can be repeated. Even stricter forms of serialism extend this rule for pitch classes to tone duration and timbre.

Cognitive Constraints on Composition

Listeners, who had internalized an expectation for unambiguous tonal centers following repeated

exposure to tonal music did not respond well to these experiments. Lerdahl (1988) argued that the widespread rejection of serial composition by listeners is not only a question of familiarity but is understandable from a cognitive perspective. To perceive and remember music, listeners need hierarchical structure. When all events are perceived on the same structural level with no one musical event subordinate to another, the music will sound “like beads on a string” with little basis to perceive, remember, and appreciate it (1988, pp. 341–342; see Dienes and Longuet-Higgins, 2004). On the other hand, listeners’ sensitivity to hierarchical structure may be limited, so its requirement as a musical property may be overstated. Listeners do not even appear to notice when music starts and ends in a different key (Cook, 1987), suggesting that short-term memory limits the extent to which hierarchical structure can be extracted. Cook (1987) proposed that theories of music that emphasize hierarchical structure are “better seen as a means of understanding the practice of tonal composers than as a means of predicting the effects of their compositions upon listeners” (p. 204).

According to Lerdahl’s (1988) view, the capacity of listeners to perceive and appreciate music is not just a question of experience and enculturation but is also driven by hard-wired perceptual and cognitive constraints. These include psychophysical limits, such as the range of perceivable frequencies, which precludes the appreciation of melodies composed with extremely high or low pitches, as well as constraints on memory. Not surprisingly, listeners find it easy to remember melodies composed with seven discrete pitches (e.g., the major scale) but difficult to remember melodies composed of 26 discrete pitch categories (as might occur in microtonal tuning). They also find it easiest to track rhythms that consist of durational contrasts that form small integer ratios (Clarke, 1999). Lerdahl (1988) proposed that qualities such as memorability and simplicity are fundamental requirements for music to be understood and appreciated. For the most part, such cognitive constraints exert an unconscious influence on composers and listeners, but composers can work consciously against such constraints in an attempt to break new ground, and listeners in search of innovative forms may actively seek out music that challenges their musical expectations.

Huron (2001a) identified several basic principles of auditory perception that are culturally encoded in the traditional pedagogy for composing with

multiple voices, known as the *rules of voice-leading*. Dating back hundreds of years and commonly taught today, these rules describe the arrangement of individual voices (e.g., soprano, alto, tenor, and bass) with respect to other voices sounding simultaneously (harmony). They include guidelines for optimal movement between successive tones within a single voice (melody), and for optimal *concurrent* motion in voices that are occurring simultaneously (counterpoint). In general, adherence to the rules of voice-leading promotes the perceptual independence of each voice. Composers such as Palestrina in the Renaissance period and J. S. Bach in the Baroque period, whose works stress differences between voices (polyphony), are remarkably consistent in their adherence to the rules of voice-leading. In contrast, for homophonic traditions such as barbershop quartet, which stress the harmonic blending of multiple voices, voice-leading rules may be deliberately disregarded.

Three selected examples illustrate the type of connections identified by Huron. First, both the dictates of harmonic writing and the distribution of pitches found in Western and non-Western music correspond to the region of frequencies associated with highest pitch *clarity*, also called the dominant pitch region (extending from F2 to C5 and centered at 300 Hz, which is approximately two semitones above middle C). This correspondence implies a tacit goal of using tones with a clear sense of pitch. Second, voice-leading rules that encourage stepwise melodic motion to a neighboring scale tone (e.g., from *mi* to *re*) reflect the mechanisms of auditory scene analysis that connect acoustic events over time into a single perceptual stream. A sequence of discrete tones varying in pitch is more likely to be perceived as a coherent whole or stream when the pitch interval between successive tones is small. Conversely, pitches that are distant from each other tend to be perceptually segregated into separate streams. Thus, to ensure the perceptual independence of voices, the pitches of successive tones within a single voice must be more proximate than the pitches of successive tones between voices. Otherwise, confusions in voice attribution may occur, particularly if voices “cross” in pitch register; a voice-leading practice that, not surprisingly, is discouraged.

Third, the spacing of tones between voices sounding simultaneously is influenced by *harmonicity*. Frequencies that fall along the harmonic series (unison, octave, perfect fifth, etc.) tend to be perceptually fused, giving rise to a single pitch sensation.

Huron (1991) found that in the polyphonic writing of Bach, these intervals are actually avoided between voices in direct proportion to the strength with which each interval promotes tonal fusion. That is, even though such intervals are consonant, they are avoided when the goal of polyphonic writing is to emphasize the independence of different voices. Tonal fusion is also promoted by positively correlated pitch motion between voices. Voice-leading rules, in turn, encourage the use of *contrary* melodic motion among voices where possible, and they discourage similar movement among voices especially if those voices are separated by a consonant interval. Indeed, parallel movement of unisons, octaves, and fifths is strictly discouraged. In short, when a composer wants the different voices of a polyphonic composition to make independent contributions to the overall musical texture, detailed steps are taken to avoid the perceptual process of tonal fusion.

Conclusions

Figure 39.1 illustrates an overview of some of the processes that underlie the acts of listening to and making music. Processes that operate automatically are differentiated from processes that rely on working memory and executive function. The two types of processes operate simultaneously. However, whereas automatic processes operate continuously, those that rely on working memory and executive function are coupled with the musical structure and influenced by the salience of unfolding events. Metric stress provides a source of salience that operates cyclically; other attributes of music (e.g., harmonically important events) also affect salience and,

hence, the amount of attention allocated. The time-varying allocation of attentional energy that arises from meter and other sources of salience is known as *rhythmic attention*.

This waxing and waning of attention accompanies many automatic processes that operate continuously and without conscious control. For example, listeners automatically internalize regularities in music, such as the frequency of occurrence of pitches (Saffran et al., 1999). In Western and non-Western music, the frequency of occurrence of pitches is correlated with the relative prominence of that pitch within a key, with the most prominent or stable tone occurring more frequently than other pitches. Composers do not consciously manipulate the frequency of occurrence of pitches. Rather, composition in Western tonal music is characterized by a set of rules for moving toward and away from the tonic. Only listeners with formal music training are aware of these rules but all listeners acquire implicit knowledge of tonality through an automatic process of statistical learning.

Listeners also respond emotionally to music using both implicit and explicit processes. Listeners automatically decode acoustic attributes associated with emotional responses, such as increases in loudness (as in a crescendo), rapid tempo, and high pitch (Ilie & Thompson, 2006, 2011). Automatic decoding of acoustic attributes is manifested in emotional interpretations of music, and it can also lead to changes in mood and arousal states. At the same time, music may trigger conscious associations: with a difficult period in one's life, a cherished memory, or an imaginary event such as an arduous journey. Such

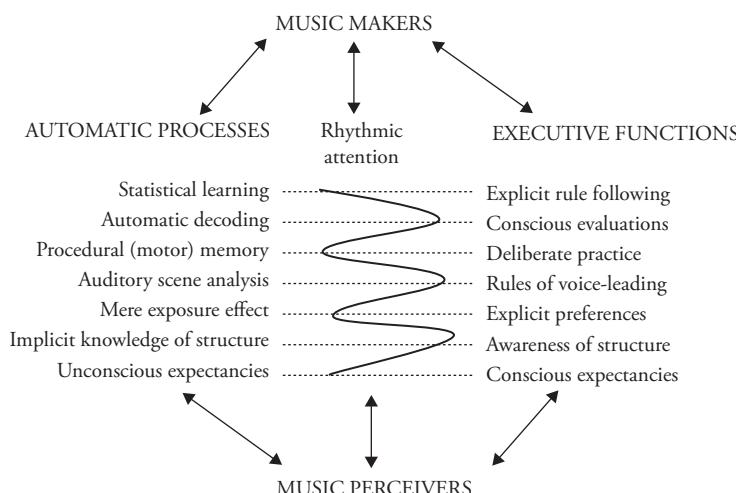


Fig. 39.1 A framework for understanding the automatic and explicit processes involved in musical thought.

associations are often part of conscious awareness and can significantly affect evaluations and experiences of music.

Such experiences are powerful, and many individuals pursue them through the craft of musicianship. After years of deliberate practice—practice that is initially effortful and consciously goal directed—elite musicians eventually encode the technical skills required for a music performance in *procedural* (motor) memory. Once encoded, performances are guided by procedural memory with little need for conscious monitoring of each motor action. The use of procedural memory to assist music performance frees up cognitive resources that can then be applied to other aspects of musical behavior, such as refining one's expressive performance or monitoring listeners' responses.

Others seek out musical experiences through the art of composition. Students of composition are initially taught *rules of voice-leading* and their compositions are initially crafted using explicit knowledge of these rules. Underlying these rules are unconscious perceptual processes that function to analyze the acoustic environment (Huron, 2001a). As musicians develop their expertise at composition, their application of the rules of voice-leading itself becomes automatized, a form of procedural memory.

Most listeners are conscious of their musical preferences, but not for how these preferences arise. Preferences can emerge automatically following passive exposure to music, reflecting implicit memory for music (the *mere exposure effect*). They may also arise through explicit routes, however, such as direct peer influence or validation of one's sense of identity.

Passive exposure to music not only leads to musical preferences; it enables listeners to build sophisticated internal representations of melodic, harmonic, tonal, and rhythmic structure. For example, the brain reliably registers subtle violations to harmony even in listeners with no musical training (reflected in EEG recordings as an early right anterior negativity or ERAN). With training, musicians acquire explicit knowledge of this structure that they can label, analyze, and critique. They may even experience musical events in terms of explicit labels such as “perfect fourth” or “dominant seventh” and their awareness of structure can affect their aesthetic experiences of music.

Knowledge of musical structure, whether implicit or explicit, provides a basis for anticipating events. Music that is conventional can be readily anticipated; music that is unconventional may violate

our expectations. Expectations provide a powerful source of emotion because they are basic to human behavior. When expectations are fulfilled, we feel a sense of stability and reward; when expectations are thwarted, we feel a sense of instability and surprise. Music makers consciously and unconsciously manipulate the expectations of listeners, creating complex and dynamic aesthetic and emotional effects as music unfolds. In most cases, listeners are unaware that violations and fulfillments of expectations are taking place, but they are usually well aware of the emotional effects of the music.

Music makers and perceivers share most of the processes that support musical thought, in spite of differences in overt behavior and expertise. In this context, it is worth noting that the distinction between listeners and musicians is a relatively recent phenomenon, promoted by the rise of the concert hall in the 19th century and escalated by the invention of the gramophone and phonograph—technologies that enabled a physical separation between music makers and music consumers. Prior to such technologies, musical behavior was mostly a group activity with no clear separation between music makers and listeners. Research into the mirror neuron system further supports a connection between musicians and listeners, in that the same neural centers may be active when perceiving or executing an action (Rizzolatti & Craighero, 2004). In particular, musical signals that arise through the actions of performers may be transformed into motor representations by perceivers (Colling & Thompson, in press; Overy & Molnar-Szakacs, 2009; Thompson & Quinto, 2011). Musically trained listeners may form detailed motor representations of these signals, but untrained listeners may also exhibit a motor response. In the absence of specialized knowledge of music production, untrained listeners can still tap to music, dance to music, perform “air guitar,” or otherwise synchronize with the music, reestablishing the shared experience between music makers and perceivers that has characterized musical behavior for much of human history.

Future Directions

The recent surge of interest in selected issues will undoubtedly lead to significant advances in the study of music and thought. First, there has been a renewed interest in evolutionary perspectives on music, following a neo-Darwinian movement across numerous disciplines (Vitouch & Ladinig, 2009; Wallin, Merker, & Brown, 2000). Adaptationist

arguments hold that the mind is specifically shaped by evolution to support musical behavior. They include the hypothesis that music originated as a sexually selected trait, like birdsong (Darwin, 2006; Miller, 2000), or from lullabies that promoted parent–infant bonds, which were crucial for infant survival (Dissanayake, 2008; Falk, 2004), or as a means of promoting group bonding and cooperation, which was essential for survival for early hominids (Huron, 2001b). Nonadaptationist accounts suggest that musical behavior is an evolutionary byproduct—a human “invention” that draws upon existing brain functions but has no survival or reproductive function in itself (Pinker, 1997). However, because musical activities can lead to various benefits for individuals during their lifetime (e.g., social bonds, cognitive benefits, parent–infant bonds, mood regulation), it has been characterized as a “transformative technology of the mind” (Patel, 2010).

A second area of renewed interest is the connection between music and emotions, which draws on data in psychology and neuroscience on the role of emotion in memory, reasoning, and problem solving (for a review, see chapters in Juslin & Sloboda, 2009). Music has a powerful capacity to communicate and induce emotions (Huron, 2006; Juslin & Västfjäll, 2008), and it appears to draw from emotional signals that are evident in speech prosody or “tone of voice” (Ilie & Thompson, 2006, 2011; Juslin & Laukka, 2003).

Finally, there has been intense scientific interest in the neuroscience of music following the rapid advances in neuroimaging techniques and other methods of understanding the brain (Patel, 2008; Peretz & Coltheart, 2003; Peretz & Zatorre, 2005). It was once believed that music was primarily the domain of the right hemisphere (Joseph, 1988). We now know that musical behaviors engage a wide range of activity across both hemispheres, and that training in music can promote structural enhancements and greater brain plasticity (Schlaug, 2001; Wan & Schlaug, 2010). Such findings highlight the enormous potential for music to be used as treatment or intervention for neurological and developmental disorders, or for cognitive impairments associated with normal aging.

References

- Aiello, R., & Williamson, A. (2002). Memory. In R. Parncutt & G. E. McPherson (Eds.), *The science and psychology of music performance* (pp. 167–181). Oxford, England: Oxford University Press.
- Ammirante, P., & Thompson, W. F. (2010). Melodic accent as ideomotor action. *Empirical Musicology Review*, 5(3), 93–106.
- Ammirante, P., Thompson, W. F., & Russo, F. A. (2011). Ideomotor effects of pitch in continuation tapping. *Quarterly Journal of Experimental Psychology*, 64(2), 381–393.
- Ashley, R. (2002). Do[n’t] change a hair for me: The art of jazz rubato. *Music Perception*, 19, 311–332.
- Bengtsson, I., & Gabrielsson, A. (1983). Analysis and synthesis of musical rhythm. In J. Sundberg (Ed.), *Studies of music performance* (pp. 27–60). Stockholm, Sweden: Royal Swedish Academy of Music.
- Boltz, M. G. (1998). Tempo discrimination of musical patterns: Effects due to pitch and rhythmic structure. *Perception and Psychophysics*, 60, 1357–1373.
- Bradley, A. (2010). *Book of rhymes: The poetics of hip hop*. New York: Basic.
- Bregman, A. S. (1990). *Auditory scene analysis*. Cambridge, MA: MIT Press.
- Chaffin, R., & Imreh, G. (2002). Practicing perfection: Piano performance as expert memory. *Psychological Science*, 13, 342–349.
- Chaffin, R., Imreh, G., & Crawford, M. (2002). *Practicing perfection: Memory and piano performance*. New York: Erlbaum.
- Clarke, E. (1988). Generative principles in music performance. In J. A. Sloboda (Ed.), *Generative processes in music: The psychology of performance* (pp. 1–26). New York: Oxford University Press.
- Clarke, E. (1999). Rhythm and timing in music. In D. Deutsch (Ed.), *Psychology of music* (pp. 473–500). San Diego: Academic Press.
- Cohen, A. J. (1991). Tonality and perception: Musical scales prompted by excerpts from the Well-Tempered Clavier of J. S. Bach. *Psychological Research*, 53, 305–314.
- Colling, L., & Thompson, W. F. (in press). Music, action, and affect. In T. Cochran, B. Fantini, & K. R. Scherer (Eds.), *The emotional power of music*. London: Oxford University Press.
- Cook, N. (1987). The perception of large-scale tonal closure. *Music Perception*, 5, 197–206.
- Cross, I. (2003a). Music and biocultural evolution. In M. Clayton, T. Hebert, & R. Middleton (Eds.), *The cultural study of music: A critical introduction* (pp. 19–30). London: Routledge.
- Cross, I. (2003b). Music and evolution: Consequences and causes. *Contemporary Music Review*, 22(3), 79–89.
- Cuddy, L. L., & Lunney, C. A. (1995). Expectancies generated by melodic intervals: Perceptual judgments of melodic continuity. *Perception and Psychophysics*, 57, 451–462.
- Darwin, C. (2006). The descent of man, and selection in relation to sex. In E.O. Wilson (Ed.), *From so simple a beginning: The four great books of Charles Darwin* (pp. 767–1248). New York: W.W. Norton (Original works published in 1871).
- Dienes, Z., & Longuet-Higgins, C. (2004). Can musical transformations be implicitly learned? *Cognitive Science*, 28, 531–558.
- Dissanayake, E. (2008). If music is the food of love, what about survival and reproductive success? *Musicae Scientiae (Special Issue)*, 169–195.
- Douglas, K. M., & Bilkey, D. K. (2007). Amusia is associated with deficits in spatial processing. *Nature Neuroscience*, 10, 915–921.
- Dunker, K. (1945). On problem solving. *Psychological Monographs*, 58, 1–110.
- Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance.

- In K. Anders Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 683–704). Cambridge, England: Cambridge University Press.
- Falk, D. (2004). Prelinguistic evolution in early hominins: Whence motherese? *Behavioral and Brain Sciences*, 27, 491–503.
- Fleming, R. (2005). *The inner voice: The making of a singer*. New York: Penguin.
- Geeves, A., Christensen, W., Sutton, J., & McIlwain, D. (2008). Review of Roger Chaffin, Gabriela Imreh, & Mary Crawford, *Practicing perfection: Memory and piano performance*. *Empirical Musicology Review*, 3(3), 163–172.
- Goebel, W., & Palmer, C. (2008). Tactile feedback and timing accuracy in piano performance. *Experimental Brain Research*, 186, 471–479.
- Goebel, W., & Palmer, C. (2009). Synchronization of timing and motion among performing musicians. *Music Perception*, 26, 427–438.
- Handel, S. (1989). *Listening*. Cambridge, MA: MIT Press.
- Helmholtz, H. L. von. (1962/1867). *Treatise on physiological optics* (Vols. 1–2). New York: Dover.
- Huron, D. (1991). Tonal consonance versus tonal fusion in polyphonic sonorities. *Music Perception*, 9, 135–154.
- Huron, D. (2001a). Tone and voice: A derivation of the rules of voice-leading from perceptual principles. *Music Perception*, 19(1), 1–64.
- Huron, D. (2001b). Is music an evolutionary adaptation? *Annals of the New York Academy of Sciences*, 930, 43–61.
- Huron, D. (2006). *Sweet expectation: Music and the psychology of expectation*. Cambridge, MA: MIT Press.
- Ilie, G. & Thompson, W. F. (2006). A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception*, 23, 319–329.
- Ilie, G., & Thompson, W. F. (2011). Experiential and cognitive changes following seven minutes exposure to music and speech. *Music Perception*, 28, 247–264.
- Jones, B., & Huang, Y. L. (1982). Space-time dependencies in psychological judgment of extent and duration: Algebraic models of the tau and kappa effects. *Psychological Bulletin*, 91(1), 128–142.
- Jones, M. R. (1989). Dynamic attending and responses to time. *Psychological Review*, 96, 459–491.
- Joseph, R. (1988). The right cerebral hemisphere: emotion, music, visual-spatial skills, body-image, dreams, and awareness. *Journal of Clinical Psychology*, 44, 630–673.
- Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129, 770–814.
- Juslin, P. N., & Sloboda, J. A. (Eds.). (2009). *Handbook of music and emotion*. Oxford, England: Oxford University Press.
- Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences*, 31, 559–575.
- Kenny, D. (2010). The role of negative emotions in performance anxiety. In P. Juslin & J. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* (pp. 425–452). New York: Oxford University Press.
- Large, E. W., & Jones, M. R. (1999). The dynamics of attending: How we track time-varying events. *Psychological Review*, 106, 119–159.
- Lehmann, A. C., & Gruber, H. (2006). Music. In K. Anders Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 457–470). Cambridge, England: Cambridge University Press.
- Lehrer, J. (2009). *The decisive moment: How the brain makes up its mind*. Melbourne, Australia: Text.
- Leman, M. (2009). *Embodied music cognition and mediation technology*. Cambridge, MA: MIT Press.
- Lerdahl, F. (1988). Cognitive constraints on compositional systems. In J. A. Sloboda (Ed.), *Generative processes in music: Psychology of performance, improvisation, and composition* (pp. 231–259). Oxford, England: Clarendon.
- Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music*. Cambridge, MA: MIT Press.
- Livingstone, S. R., Thompson, W. F., & Russo, F. A. (2009). Facial expressions and emotional singing: A study of perception and production with motion capture and electromyography. *Music Perception*, 26, 475–488.
- Maier, N. R. F. (1931). Reasoning in humans: 11. The solution of a problem and its appearance in consciousness. *Journal of Comparative Psychology*, 12, 181–194.
- Mandler, G. (1984). *Mind and body: Psychology of emotion and stress*. New York: Norton.
- McAdams, S. (2004). Problem solving strategies in music composition: A case study. *Music Perception*, 21, 391–429.
- Meyer, L. B. (1956). *Emotion and meaning in music*. Chicago, IL: University of Chicago Press.
- Miller, G. (2000). Evolution of human music through sexual selection. In N. L. Wallin, B. Merker, & S. Brown (Eds.), *The origins of music* (pp. 329–360). Cambridge, MA: MIT Press.
- Nan, Y., Sun, Y., & Peretz, I. (2010). Congenital amusia in speakers of a tone language: Association with lexical tone agnosia. *Brain*, 133, 2635–2642.
- Narmour, E. (1992). *The analysis and cognition of melodic complexity: The implication-realization model*. Chicago, IL: University of Chicago Press.
- Overy, K., & Molnar-Szakacs, I. (2009). Being together in time: Musical Experience and the mirror-neuron system. *Music Perception*, 26(5), 489–504.
- Palmer, C. (1989). Mapping musical thought to musical performance. *Journal of Experimental Psychology: Human Perception and Performance*, 15, 331–346.
- Palmer, C. (1997). Music performance. *Annual Review of Psychology*, 48, 115–138.
- Palmer, C., Koopmans, E., Loehr, J. D., & Carter, C. (2009). Movement-related feedback and temporal accuracy in clarinet performance. *Music Perception*, 26, 439–449.
- Patel, A. (2008). *Music, language, and the brain*. Oxford, England: Oxford University Press.
- Patel, A. D. (2010). Music, biological evolution, and the brain. In M. Bailar (Ed.), *Emerging disciplines* (pp. 91–144). Houston, TX: Rice University Press.
- Patel, A., Wong, M., Foxton, J., Lochy, A., & Peretz, I. (2008). Speech intonation perception deficits in musical tone deafness (congenital amusia). *Music Perception*, 25, 357–368.
- Peretz, I., & Coltheart, M. (2003). Modularity of music processing. *Nature Neuroscience*, 6, 688–691.
- Peretz, I., Gaudreau, D., & Bonnel, A.-M. (1998). Exposure effects on music preference and recognition. *Memory and Cognition*, 26, 884–902.
- Peretz, I., & Hyde, K. (2003). What is specific to music processing? Insights from congenital amusia. *Trends in Cognitive Sciences*, 7(8), 362–367.
- Peretz, I., & Zatorre, R. (2005). Brain organization for music processing. *Annual Review of Psychology*, 56, 89–114.

- Pinker, S. (1997). *How the mind works*. London: Allen Lane.
- Pressing, J. (1984). Cognitive processes in improvisation. In W. R. Crozier & A. J. Chapman (Eds.), *Cognitive processes in the perception of art* (pp. 345–363). Amsterdam, Netherlands: Elsevier Science.
- Pressing, J. (1988). Improvisation: Methods and models. In J. A. Sloboda (Ed.), *Generative processes in music: The psychology of performance, improvisation, and composition* (pp. 129–178). Oxford, England: Clarendon.
- Prince, J. B., Schmuckler, M. A., & Thompson, W. F. (2009). The effect of task and pitch structure on pitch-time interactions in music. *Memory and Cognition*, 37, 368–381.
- Quinto, L., Thompson, W. F., & Russo, F. A., & Trehub, S. (2010). The McGurk effect in singing. *Attention, Perception and Psychophysics*, 72, 1450–1454.
- Rasch, R. A. (1988). Timing and synchronization in ensemble performance. In J. A. Sloboda (Ed.), *Generative processes in music: The psychology of performance, improvisation, and composition* (pp. 70–90). Oxford, England: Clarendon Press.
- Repp, B. (1993 [1938]). Music as motion: A synopsis of Alexander Truslit's *Gestaltung und Bewegung in der Musik*. *Psychology of Music*, 21, 48–72.
- Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27, 169–192.
- Russo, F. A., & Thompson, W. F. (2005a). An interval size illusion: Extra pitch influences on the perceived size of melodic intervals. *Perception and Psychophysics*, 67(4), 559–568.
- Russo, F. A., & Thompson, W. F. (2005b). The subjective size of melodic intervals over a two-octave range. *Psychonomic Bulletin and Review*, 12, 1068–1075.
- Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, 70, 27–52.
- Schellenberg, E. G. (1997). Simplifying the implication-realization model of musical expectancy. *Music Perception*, 14(3), 295–318.
- Schellenberg, E. G., & Trehub, S. (1999). Culture-general and culture-specific factors in the discrimination of melodies. *Journal of Experimental Child Psychology*, 74, 107–127.
- Schlaug, G. (2001). The brain of musicians. A model for functional and structural adaptation. *Annals of the New York Academy of Sciences*, 930, 281–299.
- Seashore, C. E. (1938). *Psychology of music*. New York: McGraw-Hill.
- Sloboda, J. A. (2000). Individual difference in music performance. *Trends in Cognitive Science*, 4, 397–403.
- Sloboda, J. A. (2004). *Exploring the musical mind*. Oxford, England: Oxford University Press.
- Sloboda, J. A., Davidson, J. W., Howe, M. J. A., & Moore, D. C. (1996). The role of practice in the development of performing musicians. *British Journal of Psychology*, 87, 287–309.
- Thompson, W. F., Balkwill, L. L., & Vernescu, R. (2000). Expectancies generated by recent exposure to music. *Memory and Cognition*, 28, 547–555.
- Thompson, W. F., & Cuddy, L. L. (1997). Music performance and the perception of key. *Journal of Experimental Psychology: Human Perception and Performance*, 23, 116–135.
- Thompson, W. F., Cuddy, L. L., & Plaus, C. (1997). Expectancies generated by melodic intervals: Evaluation of principles of melodic implication in a melody completion task. *Perception and Psychophysics*, 59(7), 1069–1076.
- Thompson, W. F., Graham, P., & Russo, F. A. (2005). Seeing music performance: Visual influences on perception and experience. *Semiotica*, 156(1/4), 177–201.
- Thompson, W. F., & Quinto, L. (2011). Music and emotion: Psychological considerations. In P. Goldie & E. Schellekens (Eds.), *The aesthetic mind: Philosophy and psychology* (pp. 357–375). New York: Oxford University Press.
- Thompson, W. F., & Russo, F. A. (2007). Facing the music. *Psychological Science*, 18, 756–757.
- Thompson, W. F., Russo, F. A., & Livingstone, S. (2010). Facial expressions of pitch structure in music performance. *Psychonomic Bulletin and Review*, 17, 317–322.
- Thompson, W. F., Russo, F. A., & Quinto, L. (2008). Audio-visual integration of emotional cues in song. *Cognition and Emotion*, 22, 1–14.
- Thompson, W. F., & Stainton, M. (1998). Expectancy in Bohemian folk song melodies: Evaluation of implicative principles for Implicative and closural intervals. *Music Perception*, 15, 231–252.
- Trainor, L., & Trehub, S. (1992). A comparison of infants' and adults' sensitivity to Western musical structure. *Journal of Experimental Psychology: Human Perception and Performance*, 18, 394–402.
- Trainor, L., & Trehub, S. (1994). Key membership and implied harmony in Western tonal music: Developmental perspectives. *Perception and Psychophysics*, 56, 125–132.
- Trehub, S., Cohen, A., Thorpe, L., & Morrongiello, B. (1986). Development of the perception of musical relations: Semitone and diatonic structure. *Journal of Experimental Psychology: Human Perception and Performance*, 12, 295–301.
- Vitouch, O., & Ladinig, O. (Eds.). (2009). Music and evolution [Special Issue]. *Musicae Scientiae*, 2009/10.
- Wallin, N., Merker, B., & Brown, S. (Eds.). (2000). *The origins of music*. Cambridge, MA: MIT Press.
- Wan, C. Y., & Schlaug, G. (2010). Music making as a tool for promoting brain plasticity across the life span. *The Neuroscientist*, 16, 566–577.
- Wilson, W. R. (1979). Feeling more than we can know: Exposure effects without learning. *Journal of Personality and Social Psychology*, 37, 811–821.
- Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, 9, 1–27.

Learning to Think: Cognitive Mechanisms of Knowledge Transfer

Kenneth R. Koedinger and Ido Roll

Abstract

Learning to think is about transfer. The scope of transfer is essentially a knowledge representation question. Experiences during learning can lead to alternative latent representations of the acquired knowledge, not all of which are equally useful. Productive learning facilitates a general representation that yields accurate behavior in a large variety of new situations, thus enabling transfer. This chapter explores two hypotheses. First, learning to think happens in pieces and these pieces, or knowledge components, are the basis of a mechanistic explanation of transfer. This hypothesis yields an instructional engineering prescription: that scientific methods of cognitive task analysis can be used to discover these knowledge components, and the resulting cognitive models can be used to redesign instruction so as to foster better transfer. The second hypothesis is that symbolic languages act as agents of transfer by focusing learning on abstract knowledge components that can enhance thinking across a wide variety of situations. The language of algebra is a prime example and we use it to illustrate (1) that cognitive task analysis can reveal knowledge components hidden to educators; (2) that such components may be acquired, like first language grammar rules, implicitly through practice; (3) that these components may be “big ideas” not in their complexity but in terms of their usefulness as they produce transfer across contexts; and (4) that domain-specific knowledge analysis is critical to effective application of domain-general instructional strategies.

Key Words: computational modeling, language and math learning, educational technology, transfer, cognitive tutors, cognitive task analysis, *in vivo* experiments

“Learning to think” is different from “learning” in that it implies that a learner achieves an increase in more general intellectual capability, rather than just in more specific domain content. Learning to think implies more than learning English, learning math, learning history, or learning science. In other words, learning to think implies transfer (Barnett & Ceci, 2002; Gick & Holyoak, 1983; Singley & Anderson, 1989). Is it possible to “learn to think”; that is, is general transfer possible? A simple “yes or no” answer is not to be expected. A substantial amount of transfer can be achieved, especially with scientifically designed instruction, but full general

transfer is a tempting dream, not a practical reality. It is not possible to teach children how to be generally intelligent, experts in all domains, without specific instruction in some domains. But dreams can help inspire action, and the action we need is research and development to understand human learning capacities and constraints and to design and evaluate instruction that achieves *as-general-as-possible* transfer. Progress is required not only in domain-general cognitive and learning science but also in domain-specific analysis of the content domains across and within which we hope to see general transfer.

Components of Learning to Think and How Symbolic Languages Help

One way to learn to think is learning languages with which to think more powerfully. The obvious case is learning a natural language, like English. It seems uncontroversial that a good part of what makes human thinking so powerful is our capability for language (cf. Tomasello, 2000; see Gleitman & Papafragou, Chapter 28). The spoken form of natural language not only facilitates communication and collaborative thinking but also provides a medium for reasoning and logical thinking (Polk & Newell, 1995). The written form further facilitates collaborative endeavors over wide stretches of time and space and is also arguably a vehicle for improving thought. Anyone who has written a manuscript, like this one, is likely to have had the experience that the writing process changes one's thinking and yields a better product than spontaneous speech would have. Clearly, language-learning activities are a huge part of learning to think and a major responsibility of our educational systems.

But natural languages like English are not the only languages we use to enhance our thinking. By "language," in a more general sense, we mean a culturally transmitted symbolic system including any commonly used form of external representation. Examples include symbolic algebra and other mathematical notation systems (e.g., calculus notation, probability and statistics notation), Cartesian graphing and other scientific visualization techniques, and computer programming languages, including the huge and growing number of end-user programming languages (e.g., functions in Excel or html). We create external symbols (including pictures and diagrams; see Hegarty & Stull, Chapter 31) to make abstract ideas more available to our brains' powerful perceptual processing and learning mechanisms (e.g., Goldstone, Landy, & Son, 2010; Koedinger & Anderson, 1990). These forms make the abstract concrete and leverage thinking by allowing easier processing of the abstract ideas (e.g., Larkin & Simon, 1987).

In this chapter we explore two themes. The first theme is the idea that external symbol systems (languages in the broad sense) greatly enhance the power of our thinking and learning. They organize the changes internal to the mind that are necessary to implement learning and transfer (cf. Goldstone et al., 2010; Novick, 1990).

The second theme is that learning grows in "pieces" and thinking involves using those pieces in

new combinations. The amazing flexibility humans can exhibit in thinking and learning would not be possible if it were not for their extendable and reusable base of knowledge. Many pieces of knowledge have analogs or near analogs in external symbol systems, such as knowledge of English letters and words, but many do not. Such "knowledge components" (Koedinger, Corbett, & Perfetti, 2010) can include categories for which there is no word (e.g., a category of objects my 2-year-old calls "phones" but which includes remote controls and small blocks of wood). They can be "pragmatic reasoning schemas" (Cheng & Holyoak, 1985; see Evans, Chapter 8) that support correct reasoning about certain kinds of social rules, but not about abstract logical rules. They can be the "intuitions," heuristics, or abstract plans that guide our search, decision making, and discoveries. They can be metacognitive or learning strategies, like knowing to try to "self-explain" a worked example (Chi, Bassok, Lewis, Reimann, & Glaser, 1989) or to cover up one's notes and try to recall what's there while studying (cf. Pashler et al., 2007).

This knowledge component view is supported by researchers who have done detailed cognitive analysis of complex real-world learning domains, such as learning physics (diSessa, 1993; Minstrell, 2001; VanLehn, 1999), mathematics (Koedinger & Corbett, 2006), legal argumentation (Aleven, 2006), or programming (e.g., Pirolli & Anderson, 1985). The view is also supported by attempts to create large-scale computational models of complex human reasoning and problem solving (e.g., Anderson & Lebiere, 1998; Newell, 1990). With respect to transfer, this view echoes the classic "identical elements" conception of transfer (Thorndike & Woodworth, 1901) but is enhanced by advances in cognitive theory and computational modeling (cf. Singley & Anderson, 1989). The identical elements, or units of transfer, are no longer stimulus-response links but are latent representations of components of tasks or their underlying structure. It is not sufficient to provide behavioral task (stimulus-response) descriptions of such elements; rather, we need a language of abstraction for specifying cognitive representations that will often generalize over, or be applicable across, many situations. In computational modeling terms, there need to be "variables" (or some functional equivalent) in the formalism or language used (by cognitive scientists) to represent general cognitive elements (see Doumas & Hummel, Chapter 5).

In this chapter, we begin by elaborating the knowledge component view on transfer. We then provide examples of identifying and modeling knowledge components in algebra and show how these models can be tested in classroom studies. These studies have also identified the missing components—missing both in terms of students lacking the relevant knowledge, as well as cognitive scientists overlooking these pieces in their puzzle of the domain. We then argue that these missing components are intimately tied with the symbolic forms of the language itself. Next, we demonstrate how supporting language acquisition helps students achieve better transfer. Last, we discuss various aspects of applying laboratory-derived instructional principles in the classroom, and we review methods for knowledge component (or cognitive model) discovery from data. The conclusion raises open questions about the insufficiently understood interplay between language-mediated and non-language-mediated processes in transfer and learning to think.

Knowledge Components as Carriers of Transfer

Questions of transfer of learning have been addressed by a number of contemporary cognitive scientists (e.g., Barnett & Ceci, 2002; Singley & Anderson, 1989; see Bassok and Novick, Chapter 21). Much discussion has been around issues of what is “far” transfer, how much transfer instructional improvements can achieve, and what kind of instruction does so. A key observation is that when transfer occurs some change in the mind of the student is carrying that transfer from the instructional setting to the transfer setting. It might be a new skill, a general schema, a new mental model, better metacognition (see McGillivray et al., Chapter 33), better learning strategies, a change in epistemological stance, new motivation or disposition toward learning, or a change in self-beliefs about learning or social roles.¹ In the Knowledge-Learning-Instruction (KLI) Framework, Koedinger et al. (2010) use the term “knowledge component” to include all these possible variations and provide a taxonomy of kinds of knowledge components. Knowledge components are the carriers of transfer.

To better understand how knowledge components act as agents of transfer, one should identify the breadth or scope of applicability of those knowledge components in tasks, problems, or situations of interest. In other words, in how many different kinds of situations does the acquired knowledge

apply, and what are its boundaries? Understanding the answer to this question allows us to design instruction that better supports transfer. For example, instruction on computer programming might yield knowledge that applies (*a*) only to programming tasks that are quite similar to those used in instruction, (*b*) to any programming task involving the same programming language, (*c*) to programming in other languages (e.g., Singley & Anderson, 1989), or (*d*) to reasoning tasks outside of programming, like trouble-shooting or “debugging” a set of directions (e.g., Klahr & Carver, 1988).²

These two questions suggest a scientific path toward achieving more general transfer or learning to think. A key step along that path is identifying those knowledge components that are as broad or as general as possible in their scope of application. Take, for example, the algebraic task of combining like terms (e.g., $3x + 4x$). Students may learn this skill as a mental equivalent of something like “combine each number before each x.” This encoding produces correct results in some situations, but not all. It is overly specific in that it does not apply to the cases like $x + 5x$ where the coefficient of x (1) is not visually apparent. It might also yield $-3x + 4x$ as $-7x$, if the mental skill encodes “number before” too specifically as a positive number rather than more generally as a signed number (e.g., Li et al., 2010). Acquired knowledge may also produce incorrect responses by being overly general. For example, if students have encoded “number before each x” too generally, they may convert $3(x + 2) + 4x$ to $7x + 2$.

Big Ideas and Useful Ideas

In the search for transfer-enabling knowledge components, special attention has been given to the notion of “big ideas,” which has been a rallying cry in much educational reform, particularly in mathematics. It is worth reflecting on what the “big” in big idea means. It is often used in contrast with facts, procedures, or skills and associated with concepts, conceptual structures, or mental models. For instance, Schoenfeld (2007, p. 548) makes a contrast between “long lists of skills” and “big ideas,” and Papert (2000, p. 721) characterizes school as “a bias against ideas in favor of skills and facts.”

A particularly tempting example of this is the general problem-solving strategies of the sort mathematician George Polya (1957) identified in his reflections on his own mathematics thinking and teaching. Experimental efforts to investigate the effect of instruction designed to teach such general

problem-solving strategies have met with limited success. Post and Brennan (1976), for instance, found no improvement from teaching general problem-solving heuristics such as “determine what is given” and “check your results.” Schoenfeld (1985) developed more specific versions of Polya’s heuristics, and Singley and Anderson’s (1989, p. 231) analysis of that study suggests an important reason for caution in pursuing the “big idea” approach. They noted that Schoenfeld’s heuristics that led to transfer were ones that indicated when the heuristic should be applied, such as “If there is an integer parameter, look for an inductive argument.” Other heuristics, such as “Draw a diagram if at all possible,” did not indicate conditions of applicability and did not lead to transfer. Students can learn such heuristics, in the sense that they can repeat them back, but they do not get any use out of them because it is not clear when they apply. So a first caution is that sometimes an apparently as-general-as-possible knowledge component may not lead to broad transfer because it is *too vague to be useful*.

Just because a student knows general strategies for working backward, or problem decomposition, does not mean that he or she can successfully execute those strategies in a specific context. Haverty, Koedinger, Klahr, and Alibali (2000) provide an informative example. They investigated college students’ ability to induce functions, like $y = x^*(x - 1)/2$, from tables of x-y pairs, like {(2, 1) (3, 3) (4, 6) (5, 10)}. They found that all students engaged in working backward by performing operations on the y values, such as dividing each by the corresponding x value to produce {.5 1 3/2 2}. However, those who succeeded were differentiated from those that did not by recognizing that this pattern is linear (increasing by $\frac{1}{2}$). In other words, it was specific fluency in number sense that distinguished students, not general problem-solving skills that all students manifest. Thus, a second caution regarding the search for big ideas to yield far transfer is that many general concepts or strategies require the learner to obtain domain-specific knowledge in order to apply those general strategies effectively.

A third caution is that some general problem-solving or critical-thinking skills may be relatively easy to learn, in the sense that they are generally acquired without any formal schooling. Lehrer, Guckenber, and Sancilio (1988) taught third graders a general “problem decomposition heuristic” as part of an intensive 12-week curriculum surrounding the LOGO programming language. Although

they found evidence of transfer of some big ideas, they found no evidence of improvement in general problem decomposition as measured by puzzle tasks. It may be that many children acquire the relevant problem decomposition skills through prior experiences. Another example can be found in the self-explanation literature. One consistent and surprising result is that merely prompting students to self-explain improves learning, even without teaching students how to self-explain productively (Chi et al., 1989; Siegler, 2002). While improving the dispositions toward self-explanation is an important goal, it seems that the skill of knowing how to self-explain does not need much support.

The temptation to seek really big ideas is strong and these cautions are nuanced. Statements like the following are indicative: A recent National Academy of Education paper (2009) called for better assessment of “skills such as adapting one’s knowledge to answer new and unfamiliar questions.” This statement and the surrounding text, which is a call for assessments to measure “higher order, problem-solving skills,” implies that there is a general skill of “adapting one’s knowledge” that can be acquired, measured, and applied generally. It is quite unlikely, however, that there is a single, general skill of adapting knowledge. Adapting one’s knowledge is not a single skill, but, more likely, a complex set of skills that have domain-specific ties—some adaptations come easily, when the domain knowledge is in place, and others do not. There *may* be general skills that students can acquire for better adapting knowledge, but until we have identified assessments that can measure them, we should not assume that they exist.

The search for big ideas assumes that some ideas are sophisticated enough to be applied across a wide range of domains and tasks. But focusing on the complexity or sophistication of the idea itself is not sufficient if we are aiming for more effective transfer and learning to think. The issue is not the size of the idea itself, but rather the size of what the idea opens up for a learner. More precisely, it is about the size of the productive *reuse* of the idea. Some sophisticated concepts indeed get a lot of reuse, but so do some simpler facts, associations, or procedural knowledge components. For instance, knowing the phoneme associated with the letter “s” is not big in the sense of the idea being big—this fact is a small and simple one. However, it is big in the sense of its reuse. As children acquire it and the 36 or so phonemes associated with the 26 letters,³ a whole new world opens

up for them. They can use and reuse this relatively small set of knowledge components to identify (decode and pronounce) a much larger set of words in text. They can thus read words that they have not read before. Many will be words they already know from spoken language and some will be new words, which they can begin to acquire in context. Furthermore, this decoding capability (made possible by this small set of grapheme->phoneme knowledge components) greatly increases the learning capabilities of the child—he or she can now learn by reading in addition to listening, watching, and doing.⁴ Thus, rather than simply search for *big* ideas, we should be searching for *useful* ideas.

Small-big ideas are knowledge components that may be quite specific but are big in the scope of their use—they get reused in many contexts. The notion is similar to Gagné's “vertical transfer” in its emphasis on components that can combine with others to produce broader competency. However, unlike vertical transfer, small-big ideas are not limited to within-domain transfer. Small-big ideas may extend beyond their nominal domain to improve performance or learning in other domains. Phonemes are nominally in the reading domain, but acquisition of them improves learning in history, science, math, and so on—all domains in which reading is part of learning. Similarly, the distributive property is nominally in the algebra domain, but acquisition of it (and the cluster of skills associated with it) supports performance and learning in physics, chemistry, engineering, statistics, computer science, and so on—in the STEM disciplines generally.

Testing Knowledge Component Models in Instructional Contexts

Beginning in the 1980s, John Anderson and colleagues have put the series of ACT theories of cognition (e.g., Anderson, 1983) to test through the development of a brand of intelligent tutoring systems, called Cognitive Tutors (Anderson, Corbett, Koedinger, & Pelletier, 1995). Since then the work has greatly expanded in its dissemination—over 500,000 students a year use the Cognitive Tutor Algebra course (e.g., Ritter, Anderson, Koedinger, & Corbett, 2007)—and in its scope—Cognitive Tutors and variations thereof have been created for a wide variety of content, including intercultural competence (Ogan, Aleven, & Jones, 2010), statistics (Lovett, Meyer, & Thille, 2008), chemistry (Yaron et al., 2010), and genetics (Corbett, Kauffman, MacLaren, Wagner, & Jones, 2010). Many large-scale evaluations of these

courses have demonstrated their effectiveness (Ritter, Kulikowich, Lei, McGuire, & Morgan, 2007). More important for this chapter, tutoring systems (and online courses more generally) have become platforms for advancing research on thinking and learning in the field and in the context of substantial knowledge-based academic content. This makes it possible to test and advance theories of learning that may be overgeneralized or otherwise inaccurate given their origins in laboratory settings and typically knowledge-lean content. Such technology allows us to carry out well-controlled studies in the classroom environment and to collect detailed moment-by-moment data. Many of the research findings of such research can be found in the open research wiki of the Pittsburgh Science of Learning Center at <http://www.learnlab.org/research/wiki>.

A key tenet of the ACT-R theory is that human knowledge is modular—it is acquired and employed in relatively small pieces (Anderson & Lebiere, 1998). Although such pieces can be recombined in many different ways, they are not completely abstracted from context. Indeed, a second key tenet is that knowledge is context specific. A procedural form of knowledge (implicit knowledge for doing) is characterized by an if-then production rule notation (see Doumas & Hummel, Chapter 5), whereby the context in which the production applies is specified in the if-part and an associated physical action, subgoal, or knowledge retrieval request is specified in the then-part. Similarly, a declarative form of knowledge (explicit or directly accessible knowledge that can be visualized or verbalized) has retrieval characteristics that depend, in part, on the *context* of other active knowledge (the more that related knowledge is active, the more likely to retrieve the target knowledge).

These tenets lead to two important ideas for instructional design. First, it is possible to create specifically targeted instruction that isolates the learning of a particular knowledge component. Second, it is critical to design instruction so that knowledge components are acquired with appropriate context cues or features so that they generalize or transfer broadly, but accurately. Thus, isolated practice cannot be too decontextualized, otherwise inert or shallow knowledge acquisition may result.

In the process of applying the ACT intelligent tutoring systems to support learning of programming and mathematics, eight principles of instructional design were formulated to be consistent with ACT and with experience in developing, deploying,

and evaluating these systems (Anderson, Corbett, Koedinger, & Pelletier, 1995). One of these principles is that tutor design should be based on a knowledge component analysis of the target domain.⁵ This principle emphasizes the importance of the modular nature of human knowledge and the great value of domain-specific cognitive task analysis for producing effective instruction. Clark et al. (2007) describe a meta-analysis of seven studies comparing existing instruction with redesigned instruction based on a cognitive task analysis, which yielded an average effect size of 1.7 (i.e., students who received the redesigned instruction scored 1.7 standard deviations better on posttests than did students who received normal instruction).

The theory of transfer, as briefly outlined earlier, makes specific predictions about students' learning. For example, one prediction is that knowledge of a specific component can manifest itself in many different contexts. Students who acquire multiplication knowledge (with general and accurate retrieval features) can apply it to different numbers (e.g., $2 \times 3 = ?$; $6 \times 5 = ?$), to more complex symbolic problems (e.g., $2 \times 3 + 5 = ?$), and to problems presented as or emerging in a broader situation (Danny bought two pens for \$3 each). Another useful hypothesis is that performance on each component improves with practice. Cognitive Tutors and paper-based assessments allow us to put these hypotheses and analysis of the domain to test. By analyzing students' behavior on problems that share knowledge components, we can evaluate whether our analysis of knowledge components is accurate.

This idea can be illustrated with a story about algebra story problems. The cognitive science literature includes statements about how students "find word problems...more difficult to solve than problems presented in symbolic format (e.g., algebraic equations)" (Cummins et al., 1988, p. 405). When asked to predict student performance, teachers and educators indicate that algebra story problems are harder for students to solve than matched symbolic equations, since students need to first translate these word problems into symbolic notation (Nathan & Koedinger, 2000).

Koedinger and Nathan (2004) compared students' performance on story problems and matched equations, and discovered that the assumed knowledge component analysis (e.g., that equations are needed to solve story problems) was incorrect. They found that beginning algebra students are actually better able to solve introductory

story and word problems than matched equations. For instance, students were 62% correct on word problems such as, "Starting with some number, if I multiply it by 6 and then add 66, I get 81.9. What number did I start with?" but only 43% were correct on matched equations, such as " $x \times 6 + 66 = 81.90$."

One striking fact regarding these studies is the contrast between beliefs of researchers and educators on the one hand and actual student performance on the other. In this example, students' actual performance is at odds with the predictions of scientists and teachers alike. Koedinger and Nathan's (2004) domain analysis revealed previously unrecognized knowledge demands in acquiring symbolic skills. This analysis pinpoints knowledge components for symbolic language comprehension.

Learning to Think by Learning Languages to Think With

The fact that students are better at solving word problems than solving equations may sound counter to our point that learning symbolic languages, symbolic algebra in this case, facilitates thinking. However, our point is not that being "given" a symbolic language suddenly makes one a better thinker, but that the payoff for *learning* a symbolic language is more powerful thinking. That students are still struggling with the language of algebra many months into a course is surprising to many and, ironically, more so to those who have succeeded in doing so. For instance, high school algebra teachers are more likely to make the wrong prediction (equations are easier) than elementary or middle school teachers (Nathan & Koedinger, 2000). It seems that many successful algebra learners do not have good explicit memory for, and perhaps did not have much explicit awareness of, all the work they did (or their brains did) while learning algebra. This observation is consistent with the hypothesis that much (not all!) of algebra learning is done with little awareness of many of the mental changes that are taking place. Although algebra textbooks and classes include lots of verbal instruction, much of the process of learning appears to occur while students are studying examples and practicing on problems (cf. Matsuda et al., 2008; Zhu & Simon, 1987), and neither examples nor problems contain verbal descriptions of the to-be-learned patterns or rules. Many of the pattern, rule, or schema induction processes that carry out this learning are implicit (see Evans, Chapter 8), that is, are not mediated by (nor

involve the subvocalization of) the verbal rules read in the text or heard in class.⁶

Although our empirical and theoretical support for this claim comes largely from research in the algebra domain, this substantial nonverbal character of learning may extend beyond algebra. That is, much of our learning even in the academic context may be implicit (nonverbally mediated), particularly so in the STEM disciplines where symbolic notations are so common (e.g., chemical symbols, genetics notations, process diagrams). This claim is less surprising when you consider that the human brain has an amazing capability for learning language (without already having language available). Does this capability stop working once we have learned our first language, so that subsequently we only learn through language and through conscious awareness? That seems unlikely. It seems more likely that the ability to learn languages without language continues to operate even after students begin using language-mediated learning strategies. The hypothesis that nonverbal learning mechanisms are critical to learning formal symbolic “languages,” like algebra, is at least worth pursuing.

Learning a symbolic language is hard, not in the sense that it feels hard (though it might), but in the sense that it takes a long time, many months or years, to reach proficiency. Children are acquiring their first spoken language, like English, during the first 4 or 5 years of life. It typically takes a few more years to learn the written form of English, that is, reading and writing. It takes at least a year for most students to (begin to) learn algebra, and most only become fluent as they continue to use (and improve) their algebra skills in downstream STEM courses, such as Algebra II, Calculus, Chemistry, Physics, Statistics, and Programming.

Through work on Cognitive Tutor math projects, it became increasingly apparent to the first author that students’ struggles were as or more often with the specifics of the math content than with general strategies for employing it effectively in problem solving. Some targeted instruction on strategies for general problem solving may be effective, but a major challenge for students is learning specific symbolic material (e.g., vocabulary, notational tools, principles) and specific symbolic processing machinery (interpretive procedures) to fuel those general problem-solving strategies.

The Koedinger and McLaughlin (2010) study summarized later in this chapter provides evidence of grammar learning being important to progress

in algebra. Computational modeling by Li, Cohen, and Koedinger (2010) indicates that probabilistic grammar learning mechanisms are not only capable of acquiring key aspects of algebra but appear to provide a candidate answer to a mystery in expertise development. Such mechanisms provide an explanation for how learners achieve representational (or “conceptual”) changes along the path from novice to expert, accounting not only for their improved performance (accuracy and speed) but also for acquisition of deep features (e.g., Chi, Feltovich, & Glaser, 1981) and perceptual chunks (e.g., Gobet, 2005; Koedinger & Anderson, 1990). Similar learning mechanisms (Bannard, Lieven, & Tomasello, 2009) have been demonstrated to characterize children’s language acquisition.

Non-language-mediated learning mechanisms may also be a key part of learning in other STEM domains, which involve specialized symbol systems (e.g., chemistry equations, physics principles, genetics notations) and associated semantics, new vocabulary, and problem-solving processes. These *physical symbol systems*⁷ allow our powerful perceptual pattern-finding and structure-inducing learning mechanisms to operate in new abstract worlds that are designed as *visible* metaphors of hidden scientific phenomenon. Learning to see complex scientific ideas in symbolic forms allows experts to off-load long chains of abstract reasoning into physical space (on paper or computer screens). Such chains of reasoning are difficult to do in one’s head, without the external memory and perceptual processing support of symbol systems (cf. Goldstone et al., 2010). Experts are certainly capable of chains of mental reasoning without external symbolic support. This reasoning may often be done through subvocalization or subvisualization (in our “mind’s eye”), whereby (with experience) we can simulate in our minds what we might have previously done with the external support of a symbolic language (cf. Stigler, 1984).

Improving Transfer With Domain-General and Domain-Specific Approaches

So far we have focused on the role of *domain-specific* symbols and modular knowledge acquisition in facilitating transfer. However, differences in *domain-general* instructional and learning strategies (e.g., spacing practice, comparison, self-explanation) also influence transfer (cf., Koedinger et al., 2010; Pashler et al., 2007). Can such strategies be effectively and straightforwardly applied across domains?

We first illustrate how applying domain-general instructional strategies necessitates addressing the domain-specific question of what are the as-general-as-possible knowledge components. Next, we illustrate how the nature of knowledge components in a domain may change regardless of whether one instructional strategy produces more learning and transfer than another.

Finding the Right Level of Generality to Apply an Instructional Strategy

Consider the Gick and Holyoak (1983) studies that compared the effects on an analogical transfer task of different instructional strategies (see Holyoak, Chapter 13). The domain involves “convergence” tasks whereby a story lays out a problem (e.g., about the need for radiation treatment of high intensity that reaches a brain area, but without damaging the skull and tissue surrounding it), to which the solution involves a dividing of forces along multiple paths that then converge together on a target. Gick and Hoyer found that the best transfer was achieved by instruction that asked students to compare two examples (or analogs) of a general solution schema in the context of a symbolic abstraction (a diagram) representing the general solution schema. Other studies have also demonstrated the effectiveness of prompting for example comparisons (Gentner et al., 2009; Rittle-Johnson & Star, 2009), or for providing an abstraction of a general rule, pattern, or theory behind solutions (e.g., Judd, 1908; Holland, Holyoak, Nisbett, & Thagard, 1986).

How might we best map laboratory results like these onto educational practice? One challenge is determining the general schema that is the target of instruction or, to put it in more concrete terms, the scope of tasks, examples, or analogs from which to draw for use in instruction and in assessing transfer. It is necessary to have a clear definition of the target knowledge, or, in other words, the as-general-as-possible knowledge components.

Consider the goal of applying these results to the teaching of algebra symbolization, that is, translating story problems to algebraic expressions.⁸ Figure 40.1 illustrates the potential complexity of this question (data from Koedinger & McLaughlin, 2010). Is there a general schema that covers all algebra problems (or even broader, covering all math problems or all problems including convergence problems)? Or is the schema something more narrow, like all problems whose solution is a linear expression of the

form $mx + b$? Gick and Holyoak (1983) observed that the level of similarity or dissimilarity of the analogs may play an important role in how much learning and transfer occurs. Analogs with higher similarity have the advantage that students may be more likely to make a reasonable mapping between them and induce a general schema. They have the disadvantage that the schema that is induced may not be as general as it could be.

Here are two symbolization problems that are quite close analogs:

1) Sue is a plumber and gets \$30 for showing up to a job plus \$60 per hour she works. Write an expression for how much Sue makes if she works for x hours. Answer: $60x + 30$

2) Rob is an electrician. He gets paid \$50 per hour and also gets \$35 for every job. Write an expression for how much Rob makes if he works for h hours on a job. Answer: $50h + 35$

They are both in the form $Mx + N$, where M and N are small positive integers. They have some different features, including the different numbers (values for M and N) and different cover story features (e.g., “plumber” in one but “electrician” in the other). Analogs that are even closer in similarity are possible, for example, where the cover story is the same, but only the numbers change; or where the cover story changes, but the numbers do not. Analogs that are more dissimilar can also be produced, for example, by changing the type of quantities from money to distances. Structural changes are also possible, for example, introducing problems of the $Mx - N$ form (i.e., story problems with solutions like $5x - 10$). Student performance on $Mx + N$ and $Mx - N$ forms is quite similar, which suggests that such variation is not too much—does not cross into the disadvantage side of dissimilarity—and thus transfer between such problems may well be achieved. But what about more dissimilar problems?

Would it be better to include an analog with a negative slope, that is, story problems with a solution like $800 - 40x$ (31% correct)? Might that foster more generalization such that students would be more likely to transfer their learning experience not only to other positive slope problems but also to all problems of the form $mx + b$, where m and b can be positive or negative? Might we go even further to foster even greater generalization and transfer? The generalization hierarchy of problem types (potential general schemas) in Figure 40.1 illustrates

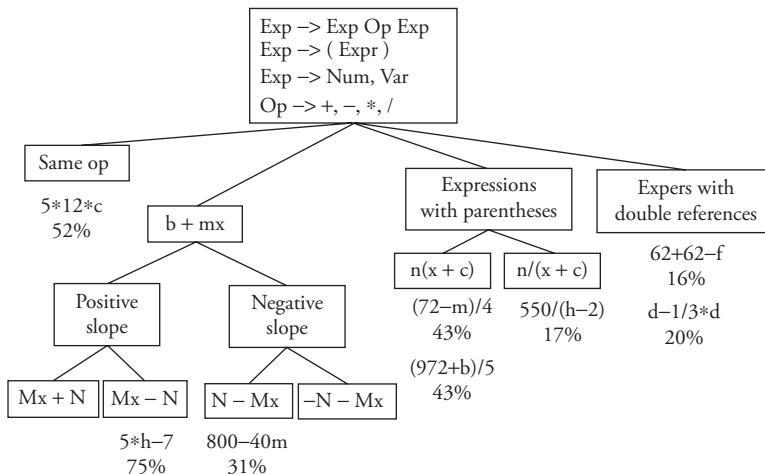


Fig. 40.1 What is (are) the general schema(s) for algebra symbolization? More broadly defined schemas (at the top) increase the potential for transfer, but more narrowly defined schemas (at the bottom) increase the probability that a general schema is induced and some transfer is achieved.

a challenge for instructional designers: Even if the instructional principle is clear and well documented, applying it to a domain of interest is not obvious. As shown next, the challenge is greater in cases where different instructional principles suggest competing paths to learning.

Knowledge-Based Dependencies in Applying Instructional Principles

Two competing instructional principles are *learning by doing* and *learning from worked-out examples*. Learning by doing (e.g., Dewey, 1916) essentially suggests that ideas or skills that we are told or shown do not stick, are not robustly learned, unless we use them. A version of this principle is another one of the Cognitive Tutor principles: Provide instruction in a problem-solving context (Anderson et al., 1995). It is related to the “testing effect” (e.g., Roediger & Karpicke, 2006), in that problem solving requires or “tests” recall in targeted contexts and so strengthens the mental link between context and appropriate action.

However, cognitive load theory (e.g., Sweller, 1988; Van Merriënboer & Sweller, 2005) suggests that premature problem-solving practice (e.g., before enough study of worked examples) produces *extraneous* cognitive load. One unfortunate outcome of intuitive instructional design is that it often introduces activities requiring more cognitive processing than necessary, which distracts students from processes relevant to learning (cf. Clark & Mayer, 2003). Indeed, many studies in science, math, and technical domains have demonstrated the “worked example effect,” whereby replacing many

problems with worked examples enhances learning (see reviews by Renkl & Atkinson, 2010; Pashler et al., 2007). In an apparent contrast with the “testing effect,” the worked example principle implies more study (of examples) and less testing (fewer problems to solve). Both principles reject the extremes of all learn by being told or all learn by doing. However, it is unclear whether the greater student assistance recommended by the worked example principle is actually, or just apparently, contradictory to the greater student challenge recommended by the testing effect.

One possible account of the apparent contradiction focuses on differences in the knowledge content involved in the corresponding instructional experiments. Experiments on the testing effect have targeted specific facts (e.g., in language learning), simpler knowledge components, and the corresponding learning theory emphasizes memory processes. In contrast, experiments on the worked example effect have targeted general schemas (e.g., in math and science learning), more complex knowledge components, and the corresponding learning theory emphasizes the induction of schemas (see Holyoak, Chapter 13). Koedinger et al.’s (2010) KLI Framework provides a principled distinction between simpler and more complex knowledge components. Schema induction may be more optimally supported by increased study of examples before turning to practice (e.g., Gentner et al., 2009; Gick & Holyoak, 1983), whereas fact memory may benefit from a faster transition from study to recall practice. However, as far as we know, no one has attempted to test this content-by-principle

interaction hypothesis, namely, that a higher ratio of study/examples to test/problems is appropriate for more complex knowledge (schema induction), whereas a lower ratio is appropriate for simpler knowledge (fact memory).

Translating general instructional approaches, such as comparison of examples, to real instructional problems may seem relatively straightforward, but this discussion suggests that it requires the answer to what can be a difficult question: What is the target schema that should guide which analogs are selected, how general should it be, and, correspondingly, how much transfer can be achieved? Typically, instructional designers make a decision based on intuition, but a more scientific approach is possible. The next section provides an example of such an approach.

Discovering a Small-Big Idea: Algebra as Language Learning

Koedinger and McLaughlin (2010) performed an experiment targeting the highest level in Figure 40.1. They identified a general schema (or knowledge component) common to all problems in this broad category. Prior empirical cognitive task analysis (Heffernan & Koedinger, 1998) had demonstrated that students' difficulties in translating story problems to algebra were not so much in understanding the English of the story but primarily in producing algebraic expressions. We hypothesized that a general knowledge component that many students were missing was (probably implicit) knowledge of the grammar of algebraic expressions involving more than one operator (see the top box in Fig. 40.1). Such students can accurately produce algebraic expressions of the form "number operator number" (e.g., $800 - y$ or $40x$), but have trouble producing expressions involving a subexpression like "number operator expression" (e.g., $800 - 40x$), that is, learning recursive grammatical patterns. We hypothesized that we could support students in learning of such recursive grammar patterns through exercises isolating the production of two-operator expressions, namely, substitution problems like "Substitute $40x$ for y in $800 - y$." We found that tutored practice on such substitution problems led to greater transfer to performance on translating two-operator stories than did tutored practice on translating one-operator stories (Koedinger & McLaughlin, 2010).

We did not provide any direct instruction on the algebraic grammar, but nevertheless students

improved in their algebraic language production. The transfer observed is consistent with the hypothesis that implicit (non-language-based) symbolic language-learning mechanisms are operative even for algebra students.

While the difference in transfer was statistically reliable, it was not large. The large differences in error rates for problems in Figure 40.1 suggest that learning recursive grammar rules is not be the only challenge for students. Problems whose solutions include parentheses are likely harder than otherwise similar problems. In fact, producing the associated expressions requires new grammar rules. It also appears that additional knowledge is needed to address problems in which a quantity serves in multiple roles, like the " d " in " $d - 1/3d$." These observations about problem difficulty suggest hypotheses for alternative instructional design. Students may benefit, for instance, from more focused instruction on the use of parentheses in algebraic expressions.

The earlier example is illustrative of using student performance data to drive cognitive task analysis, search for as-general-as-possible knowledge components, and correspondingly improve instructional design. The next section describes some strategies for fostering the process of discovering such general knowledge components.

Methods for Discovering As-General-As-Possible Elements of Transfer

We review a number of empirical methodologies for discovering transfer-enabling knowledge components or, in different terms, to perform cognitive task analysis to aid designing effective instruction for transfer. Because so much of learning is not language based and not available to our intuitions, we need techniques that bring theory and data to bear on questions of what constitute ideal instructional objectives, what are the big ideas instruction should target, and how general instructional principles can be best applied in specific content domains to produce transfer.

THINK ALOUD: EMPIRICAL ANALYSIS OF EXPERTS AND NOVICES

Having experts or novices think aloud as they solve tasks in a target domain (Ericsson & Simon, 1984) is a powerful tool for aiding in identifying the knowledge they employ. Chi et al. (1989) used think-aloud methods with physics learners to identify a potential "big-big" idea, *self-explanation*. Designing instruction that prompts students to self-

explain has been demonstrated to greatly enhance student learning in a variety of domains (see recommendation #7 in Pashler et al., 2007). As mentioned earlier, Haverty et al. (2000) performed a think-aloud procedure with students engaged in inductive pattern discovery and were surprised to find that success was differentiated not by big-big ideas like general problem-solving strategies, but by small-big ideas that produce fluency with number patterns.

One too-rarely-employed strategy for identifying as-general-as-possible elements of transfer is to use the think-aloud method on tasks for which experts lack domain-specific knowledge. For example, Schunn and Anderson (1999) asked *social* psychologists to design experiments to address a *cognitive* psychology question so as to isolate domain-general scientific inquiry skills from domain-specific experience.

DIFFICULTY FACTORS ASSESSMENT: EXPERIMENTAL ANALYSIS OF TASK FACTORS THAT REDUCE LEARNERS' PERFORMANCE

Wanting to preserve the value of getting empirical data on student task performance yet reduce the costs of think-aloud data collection and analysis, the first author began a strategy of placing item-based experimental designs into classroom quizzes. We have called this approach Difficulty Factors Assessment (DFA), and the Koedinger and Nathan (2004) story problem data described earlier is an example of this approach. A DFA is a factorial design of matched tasks or problems that vary in a multidimensional matrix of factors, for instance, whether the problem is presented in a story, in words, or in an equation, whether the unknown is in the result ($4 * 25 + 10 = x$) or start ($x * 25 + 10 = 110$) position, or whether the numbers involved are whole numbers or rational numbers. These items are distributed on multiple forms and administered to students as a quiz. We have run DFA studies in many domains, including algebra problem solving, algebra symbolization, negative numbers, fractions, data display interpretation and production, and finding areas (e.g., Heffernan & Koedinger, 1998; Koedinger & Nathan, 2004). Baker, Corbett, and Koedinger (2007) discuss how DFA studies can be used in instructional design. The idea, described earlier, of using algebraic expression substitution exercises to improve transfer in algebra symbolization, was discovered from DFA studies. In general, identifying the task factors that cause students the most

difficulty supports the instructional designer both in focusing their effort on the greatest need, and in testing whether purported forms of instructional assistance actually improve student performance.

EDUCATIONAL DATA MINING: DISCOVERING COGNITIVE MODELS FROM STUDENT PERFORMANCE AND LEARNING DATA

As we have fielded more interactive tutoring systems, they have increasingly become valuable sources of data to understand student learning (Cen, Koedinger, & Junker, 2006; Koedinger et al., 2010). In these tutoring systems, students typically solve a series of problems and the system evaluates student performance on a step-by-step basis such that error rate (did they get that step right on their own on the first try) can be logged for each task (each step in each problem). As with DFA studies, such student error data can be used to develop better cognitive models of the factors of problems or tasks that cause students difficulty. One current disadvantage of using tutor data relative to DFAs is that the set of problems given to students in the former type of study are typically not as systematically designed and administered, with matched sets of problems in a Latin square design, as they are in DFA studies.

However, there are also multiple advantages of tutor data over DFA data. First, fielded tutors allow for much more data to be naturally collected as a part of normal system use, that is, without the need to administer a special-purpose paper-based quiz. Second, student performance is automatically graded. Third, the data are more fine grained: whether the student got each step right, rather than just whether the student got the whole problem right. Even if students show their work on paper, if an error is made on an early step in a problem, the rest of the steps are typically absent or suspect. Tutoring systems, on the other hand, provide data on every step because they give students assistance so that early steps are eventually performed correctly and thus the student can attempt every step on his or her own. The fourth, and most important, advantage of tutoring systems over DFAs is that tutor data are longitudinal, providing an indication of changes in student performance over time. Seeing change in performance over time (sequence data) is of critical importance for understanding transfer.

Let us illustrate this point by contrasting how a question of knowledge decomposition (what are the elements of transfer) can sometimes be better

addressed using sequence data (from tutor log data) rather than factorial design data (from a DFA). The question in the abstract is whether there is transfer between two tasks that have some core similarity, that is, they share a deep structure (or key aspects of a deep structure) but have some significant dissimilarity, which may be either or both substantial (but solution-irrelevant) surface differences or nonshared aspects of the deep structure. Call these tasks “dissimilar analogs.” The knowledge component question concerns the components that are in common in these dissimilar analogs and what components, if any, are specific to each analog.

Consider task A and task B as dissimilar analogs where the average success rate on A is 43% and B is 27%. Table 40.1 shows an example of two such tasks.

One of the simplest knowledge component models to characterize this situation is that the harder task B requires two knowledge components, say K1 and K2, whereas the easier task A, requires just K1. This “overlap” knowledge component model predicts transfer between instruction on one of these tasks and performance on another. This situation is illustrated in Table 40.1 where task A (in the first row) is modeled by a single knowledge component representing algebra grammar knowledge for producing a recursive expression (an expression, like “ $72 - m$,” inside another expression, like “ $(72 - m)/4$ ”). Task B is represented by two knowledge components, one for comprehending English sentences and translating them to math operations, like $72 - m$ and $x/4$, and the other is the overlap, the knowledge for producing a recursive expression.

A competing *nonoverlap* model corresponds to the idea that these dissimilar analogs are not (functionally) analogs at all, but they are different topics and draw on different skills or concepts. Each task involves a separate knowledge component, say Ka for task A and Kb for task B. This model predicts no transfer. How can we use data to determine which is the correct model?

According to the “identical knowledge components” theory of transfer, the overlap knowledge component model predicts that successful instruction (e.g., tutored practice) on task A will improve performance on task B, whereas the nonoverlap knowledge component model predicts no improvement. We can distinguish these models if we have sequence data that provide performance on task B after task A for some students and before task A for others. If performance on task B is better after task A than before it, then the overlap knowledge component model is the better model. For the tasks shown in Table 40.1, students who saw Task B after seeing two substitution problems (isomorphic to Task A) indeed achieved reliably greater success, 38% correct, as compared to 27% for students who had not seen substitution problems.⁹

More generally, computer-collected student performance and learning data have been used to evaluate cognitive models and to select among alternative models (e.g., Anderson, 1993; Ohlsson & Mitrovic, 2007). Automated methods have been developed to search for a best-fitting cognitive model either purely from performance data, collected at a single time (e.g., Falmagne, Koppen, Villano, Doignon, & Johannessen, 1990), or from learning data, collected across multiple times (e.g., Cen et al., 2006).

Table 40.1 Two Dissimilar “Analogs” With Different Problems but Similar Solutions

Problem	Solution	% Correct	Knowledge Components Needed
A Substitute $72 - m$ for d in $d/4$. Write the resulting expression.	$(72 - m)/4$	43%	RecExprProd ^a
B Ann is in a rowboat on a lake. She is 800 yards from the dock. She then rows for m minutes back toward the dock. Ann rows at a speed of 40 yards per minute. Write an expression for Ann’s distance from the dock.	800 – 40x	27%	EngToMathOps ^a + RecExprProd

^aEngToMathOps = English comprehension and translation to math operations, such as $40x$ and $800 - y$. RecExprProd = Recursive expression production, like $800 - 40x$ from $40x$ and $800 - y$.

We should also be alert for decomposition opportunities, where the transfer may not be at the whole problem level (problem schemas), but at the level of intermediate steps (step schemas or knowledge components). The concept of a problem schema, which is emphasized in most studies both in the psychology (e.g., Gentner et al., 2009; Gick & Holyoak, 1983) and educational psychology (e.g., Sweller & Cooper, 1985) literature, mostly ignores the important componential character of the *construction* of human intelligence. Transfer of knowledge to novel situations comes as much or more from the reconfiguration or recombination of smaller pieces of knowledge into new wholes than from the analogical application of larger pieces of knowledge.

Conclusions and Future Directions *Learning With and Without Language*

We have focused on an aspect of learning to think that involves the leverage of symbolic systems or languages to enhance thinking and learning (see Gleitman & Papafragou, Chapter 28). This kind of learning to think is rather indirect because before one can make use of a new language to aid thinking, one must first learn that language. We have construed learning languages broadly to include learning symbolic patterns (syntax, perceptual chunks) and semantic interpretations of those patterns, as is done in many STEM disciplines. A child learning his or her first natural language is the prime example of the fact that the human mind can learn language without knowing any language to support that learning. Let us call the learning processes used in this case *non-language-mediated (NLM) learning*.¹⁰ NLM learning is responsible not only for language acquisition but also for learning of other knowledge components (e.g., How do you know how much salt to put on your food?). Such processes include perceptual learning (e.g., Gobet, 2005), unsupervised statistical learning (e.g., Blum & Mitchell, 1998), and some supervised learning in which the learner imitates or induces from correct examples (Gentner et al., 2009; Hummel & Holyoak, 2003) or gets nonverbal negative feedback from incorrect actions (e.g., Matsuda et al., 2008; VanLehn, 1987).

On the other hand, it seems clear enough, given how much instructors talk, that language is an important part of the academic learning process (cf., Michaels, O'Connor, & Resnick, 2008). Let us call the learning processes used in this case *language-mediated (LM) learning*.

Here are some related questions for future cognitive science research on learning and thinking:

- 1) *NLM learning*: What are the learning mechanisms that humans use to learn a language (at least their first language) without using language? What are the mechanisms behind learning by watching, by doing after watching, or after nonverbal feedback?
- 2) *LM learning*: What are the learning mechanisms that humans use to make use of language in learning other things? What are the mechanisms of learning by listening, by reading, by writing, by doing after verbal explanation, or after verbal feedback?
- 3) *Continued use of NLM learning*: How do NLM learning mechanisms continue to be used by learners after they have acquired the relevant language?
- 4) *Separate or subservient*: Do NLM and LM processes operate independently or do LM learning processes work by calling upon NLM learning processes?

LANGUAGE IMPROVES THINKING

We have evidence for the “language improves thinking” claim in the domain of algebra (Koedinger et al., 2008). We see that students who have learned the language of algebra are much more likely to solve a particular class of complex story problems (which do not absolutely require equations) than students who have not learned the language of algebra.¹¹

Some research has shown that prompting students to think with a symbolic language can enhance learning (Roll, Aleven, & Koedinger, 2009; Schwartz, Martin, & Pfaffman, 2005). In both of these studies, students who were asked to reason using mathematical symbols acquired a more general, transferable representation knowledge than students who were not instructed to use mathematical notations. Other studies have shown that prompting students to explain their reasoning (in English) as they solve problems or study examples helps them acquire deeper understanding of the target knowledge components (Aleven & Koedinger, 2002; Chi, De Leeuw, Chiu, & LaVancher, 1994).

Experimental evidence that language improves thinking has been collected in other domains as well. For example, 3-year-old children given instruction on relational language (e.g., *big*, *little*, *tiny*) make better abstract inferences than children without such symbolic support (Gentner, 2003). Beyond algebra,

others have also argued that human-invented symbol systems, like computer modeling languages, are “literally languages and accordingly offer new cognitive tools” (Goldstone & Wilensky, 2008).

Language-Mediated Versus Non-Language-Mediated Learning, Expert Blind Spot, and Effective Educational Design

A key point of this chapter is that too much instructional design is suboptimal because it is driven by memories of LM learning experiences. It is not driven by memories of NLM learning processes because these processes and the resulting tacit changes in knowledge are hard for learners to reflect upon. In essence, because of NLM learning processes, experts have worked harder and know more than they realize. Indeed, experts often have difficulty describing what they know (e.g., Biederman & Shiffrar, 1987). As illustrated earlier, instructors and educators can have *expert blind spots* whereby their own expertise may lead them to overestimate students’ abilities with the normative problem-solving strategies (e.g., use of algebraic equations). In general, our intuitions about what and how to teach are underinformed. Thus, cognitive task analysis can provide a powerful tool for improving our intuitions and producing more effective instruction (cf. Clark et al., 2007).

More effective educational practice could be achieved through a better understanding of the role of NLM learning in academic learning. The popular (and appropriate) rejection of the “transmission model” of instruction is a step in the right direction. Students usually do not learn simply by being told. Unfortunately, the constructivist alternative (cf., Tobias & Duffy, 2009) is sometimes *misinterpreted* to essentially mean students must teach themselves and instructors need not teach at all! A more sophisticated interpretation suggests that it is the students who should be primarily doing the talking, and the teachers’ role is to get them talking (Michaels, O’Connor, & Resnick, 2008). While there is much merit in this idea, it is driven by intuitions that LM learning is where all the action is. The merits of classroom dialog can be bolstered by a complementary emphasis on the role of NLM learning (example induction and repeated practice with feedback) and its interplay with LM learning.

Knowledge Representation and Transfer

SMALL-BIG IDEAS

The focus on big ideas tends to ignore the importance of *small-big* ideas, that is, *specific* facts,

skills, or concepts that have a very high frequency of reuse. Furthermore, the learning of big-big ideas is often mediated by the learning of external symbols—new ideas are often associated with new vocabulary. For example, new mathematical ideas like the distributive property have concise symbolic descriptions, and learning the syntax of those symbolic descriptions strengthens and may even seed the semantics of the underlying idea. How can we identify the small-big and big-big ideas that are most valuable to pursue toward improving education?

SEARCH FOR APPROPRIATE GRAIN SIZE OF TRANSFER

A related issue is the instructional decision of what level of analog to choose to target. This decision is a nontrivial part of testing the generality of basic cognitive science research in educational contexts. Can we develop empirical and theoretical approaches to identify the ideal level of generality that instructional design should target to best enhance learning and transfer?

DOES UNDERSTANDING (VIA LANGUAGE-MEDIATED LEARNING) OCCUR BEFORE OR AFTER PRACTICE (NON-LANGUAGE-MEDIATED LEARNING)?

Is understanding necessary for transfer, that is, does understanding lead transfer? Or is understanding emergent from transfer, that is, does “understanding follow transfer”? To the extent that much of learning to think is about learning specialized languages, it may be that what it means to “understand” is to develop a “metalanguage” (e.g., words like “term” and “coefficient” in algebra or “conjugation” in second language learning) that one can use to describe and reflect on the language that has been learned, as well as give words and symbols to knowledge components that were acquired via MLM learning mechanisms. This kind of understanding, that is, the development of such metalanguage, may as often be a consequence of NLM learning, rather than a source of it. Here the idea/skill is first acquired in nonverbal form (e.g., students who have learned to remove a “coefficient” in an algebra equation but don’t know the term), and later the student may learn the language to describe what he or she learned. However, other examples suggest the reverse route. For example, asking students to reason with data prior to giving them instruction may facilitate mental representations that support the

subsequent acquisition of procedural competencies (Roll, Aleven, & Koedinger, 2009; 2011; Schwartz & Martin, 2004; Schwartz, Martin, & Pfaffman, 2005).

A Literally Physical Symbol System

Thinking of these pieces of mental function or “knowledge components” as “symbols,” as proposed in Newell and Simon’s (1976) physical symbol system hypothesis, is both helpful and potentially misleading. Cognitive scientists can certainly describe these pieces in symbols, whether we use English or a computational modeling language. However, the cognitive scientist’s symbolic representation of a piece of mental function, like the code that implements a neural network model, is not the mental function itself. Whether the mental pieces *are* symbols we will leave to others (e.g., see Nilsson, 2007). The important claim here is that representing mental pieces or knowledge components in symbolic form, in some language, facilitates thinking on the part of the cognitive scientist or analyst.

Closing

The notion of cultural transmission is suggestive of language, but, while quite powerful, language is not the sole contributor to learning to think. The human brain is built on an animal brain that has perceptually based forms of learning that are effective despite functioning without the use of language (see Penn & Povinelli, Chapter 27). These learning mechanisms, we have argued, are still in use by adult learners with language abilities. It is an interesting and important scientific goal to either disprove this hypothesis or to better identify and understand these NLM learning processes and how they interact with LM learning. Such an endeavor will also have important practical consequences for educational improvement.

Notes

1. That such changes are not directly observable poses a challenge to scientific advance. One way to gain leverage is to create scientific languages to describe these hidden changes, as has been frequently done, for instance, for genes, elements in chemical compounds, or atoms. That is why computational modeling is so important to advance cognitive science and fuel better educational applications.

2. Or, for another example, reading instruction could target learning to decode (sound out) a particular word list (e.g., cat, dog, run, etc.) and the scope of that knowledge, if acquired, might reasonably be reading of those words in the context of

the many sentences in which they might appear. Or, in contrast, reading instruction might use related words (e.g., cat, fat, car, far, etc.) to target particular letter-to-sound mappings (e.g., c, f, a, t, and r), and the scope of that knowledge, if acquired, might reasonably be the reading of the many words that involve those letter-to-sound mappings in the many sentences in which they might appear.

3. Estimates of the number of phonemes vary. See <http://www.spellingsociety.org/journals/j30/number.php#top>.

4. Note that decoding is necessary but not sufficient for comprehension. Good decoders may not be good comprehenders, but bad decoders are bad comprehenders. Furthermore, children have a head start on comprehension through their spoken language experience and decoding opens doors to greater conceptual and vocabulary acquisition that can expand comprehension.

5. The original principle called for a “production rule” analysis, which focuses, in ACT-R terms, on procedural knowledge. We generalize to “knowledge component” analysis (Koedinger, et al., 2010) to include declarative knowledge and to account for declarative as well as procedural transfer (cf., Singley & Anderson, 1989).

6. The point is not that algebra students are unaware of learning algebra in the whole or in the strong sense of implicit learning used, for instance, in paradigms like Reber (1967) grammar tasks. Instead, we mean that there are many details being learned (like the grammar of algebra equations) that are not the result of verbally mediated reasoning. Students have difficulty explaining many of these details and, especially, how they came to know them.

7. We use “physical” here not in the sense of Newell and Simon (1976) of having a physical existence in human minds, but in the more literal sense that the symbols exist, on paper, white boards, computer screens, and so on, in perceptually available spaces.

8. Algebra symbolization is a particularly important area of algebra learning. Today computers can solve algebraic equations and such capabilities are increasingly available even on mobile phones. However, translating the semantics of a problem situation into abstract symbolic form will remain a task for humans for quite a while.

9. This statistically reliable difference ($\chi^2(1, N = 303) = 4.68$, $p = .03$) comes from alternative analysis of data in Koedinger and McLaughlin (2010).

10. Why have we used “non-language-mediated learning” instead of “implicit learning”? Implicit learning is learning that one is not aware of, but it may be possible that one is sometimes aware of the results of NLM learning and thus such learning would not qualify, at least empirically, as implicit learning. It is straightforward to tell when language was not used (by a teacher or in instructional materials) as part of instruction. That’s clearly NLM *instruction*. Pinpointing NLM *learning* is harder as subvocalization and self-explanation may count as LM learning. One form of evidence is when we can create computational models that can learn from examples without being given any verbal instructions and that behaviors of the model match those of human learners (e.g., Matsuda, Lee, Cohen, & Koedinger, 2009).

11. Data from Study 2 in Koedinger et al. (2008) show that when a student correctly solved a complex equation, the student was 82% correct on the matched complex story problem, but when the student failed on the complex equation, he or she was only 44% correct on the matched story problem. That is, symbolic competence enhances reasoning.

References

- Aleven, V. (2006). An intelligent learning environment for case-based argumentation. *Technology, Instruction, Cognition, and Learning*, 4(2), 191–241.
- Aleven, V., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science*, 26(2), 147–179.
- Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.
- Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.
- Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, 4(2), 167–207.
- Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.
- Baker, R. S. J. d., Corbett, A. T., & Koedinger, K. R. (2007). The difficulty factors approach to the design of lessons in intelligent tutor curricula. *International Journal of Artificial Intelligence in Education*, 17(4), 341–369.
- Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, 106(41), 17284–17289.
- Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, 128(4), 612–637.
- Biederman, I., & Shiffrar, M. M. (1987). Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual learning task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(4), 640–645.
- Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of Eleventh Annual Conference on Computational Learning Theory (COLT)* (pp. 92–100). New York: ACM Press.
- Cen, H., Koedinger, K. R., & Junker, B. (2006). Learning factors analysis: A general method for cognitive model evaluation and improvement. In M. Ikeda, K. D. Ashley, T-W. Chan (Eds.), *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (pp. 164–175). Berlin, Germany: Springer-Verlag.
- Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, 17, 391–416.
- Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13, 145–182.
- Chi, M. T. H., de Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3), 439–477.
- Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121–152.
- Clark, R. E., Feldon, D., van Merriënboer, J., Yates, K., & Early, S. (2007). Cognitive task analysis. In J. M. Spector, M. D. Merrill, J. J. G. van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 577–593). Mahwah, NJ: Erlbaum.
- Clark, R. C., & Mayer, R. E. (2003). *e-Learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning*. San Francisco, CA: Jossey-Bass.
- Corbett, A., Kauffman, L., MacLaren, B., Wagner, A., & Jones, E. (2010). A cognitive tutor for genetics problem solving: Learning gains and student modeling. *Journal of Educational Computing Research*, 42(2), 219–239.
- Cummins, D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology*, 20, 405–438.
- Dewey, J. (1916). *Democracy and education*. New York: Macmillan.
- diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction*, 10(2 & 3), 105–225.
- Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: The MIT Press.
- Falmagne, J-C., Koppen, M., Villano, M., Doignon, J-P., & Johannesen, L. (1990). Introduction to knowledge spaces: How to build, test, and search them. *Psychological Review*, 97, 201–224.
- Gentner, D. (2003) Why we're so smart. In D. Gentner, & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 195–236). Cambridge, MA: MIT Press.
- Gentner, D., Loewenstein, J., Thompson, L., & Forbus, K. D. (2009). Reviving inert knowledge: Analogical abstraction supports relational retrieval of past events. *Cognitive Science*, 33, 1343–1382.
- Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, 15, 1–38.
- Gobet, F. (2005). Chunking models of expertise: Implications for education. *Applied Cognitive Psychology*, 19, 183–204.
- Goldstone, R. L., Landy, D. H., & Son, J. Y. (2010). The education of perception. *Topics of Cognitive Science*, 2(2), 265–284.
- Goldstone, R. L., & Wilensky, U. (2008). Promoting transfer by grounding complex systems principles. *Journal of the Learning Sciences*, 17(4), 465–516.
- Haverty, L. A., Koedinger, K. R., Klahr, D., & Alibali, M. W. (2000). Solving induction problems in mathematics: Not-so-trivial pursuit. *Cognitive Science*, 24(2), 249–298.
- Heffernan, N., & Koedinger, K. R. (1998). A developmental model for algebra symbolization: The results of a difficulty factors assessment. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 484–489). Mahwah, NJ: Erlbaum.
- Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press.
- Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, 110, 220–264.
- Judd, C. H. (1908). The relation of special training to general intelligence. *Educational Review*, 36, 28–42.
- Koedinger, K. R., Alibali, M. W., & Nathan, M. M. (2008). Trade-offs between grounded and abstract representations: Evidence from algebra problem solving. *Cognitive Science*, 32(2), 366–397.
- Koedinger, K. R., & Anderson, J. R. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. *Cognitive Science*, 14, 511–550.
- Koedinger, K. R., & Corbett, A. T. (2006). Cognitive tutors: Technology bringing learning science to the classroom. In K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 61–78). New York: Cambridge University Press.

- Koedinger, K. R., Corbett, A. C., & Perfetti, C. (2010). The Knowledge-Learning-Instruction (KLI) framework: Toward bridging the science-practice chasm to enhance robust student learning. CMU-HCII Tech Report 10–102. Accessible via <http://reports-archive.adm.cs.cmu.edu/hcii.html>.
- Koedinger, K. R., & McLaughlin, E. A. (2010). Seeing language learning inside the math: Cognitive analysis yields transfer. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 471–476). Austin, TX: Cognitive Science Society.
- Koedinger, K. R., & Nathan, M. J. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *The Journal of the Learning Sciences*, 13(2), 129–164.
- Klahr, D., & Carver, S. M. (1988). Cognitive objectives in a LOGO debugging curriculum: Instruction, learning, and transfer. *Cognitive Psychology*, 20, 362–404.
- Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11(1), 65–99.
- Lehrer, R., Guckenber, T., & Sancilio, L. (1988). Influences of LOGO on children's intellectual development. In R. E. Mayer (Ed.), *Teaching and learning computer programming: Multiple research perspectives* (pp. 75–110). Hillsdale, NJ: Erlbaum.
- Li, N., Cohen, W. W., & Koedinger, K. R. (2010). A computational model of accelerated future learning through feature recognition. In V. Aleven, J. Kay & J. Mostow (Eds.), *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 368–370). Heidelberg, Berlin: Springer.
- Lovett, M., Meyer, O., & Thille, C. (2008). JIME - The Open Learning Initiative: Measuring the effectiveness of the OLI statistics course in accelerating student learning. *Journal Of Interactive Media In Education*, 2008(1). Retrieved February 16, 2011, from <http://jime.open.ac.uk/2008/14>
- Matsuda, N., Cohen, W., Sewall, J., Lacerda, G., & Koedinger, K. R. (2008). Why tutored problem solving may be better than example study: Theoretical implications from a simulated-student study. In A. Esma and B. Woolf (Eds.), *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (pp. 111–121), Berlin: Springer.
- Matsuda, N., Lee, A., Cohen, W. W., & Koedinger, K. R. (2009). A computational model of how learner errors arise from weak prior knowledge. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1288–1293). Cognitive Science Society, Amsterdam, Netherlands.
- Michaels, S., O'Connor, C., & Resnick, L. B. (2008). Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in the Philosophy of Education*, 27(4), 283–297.
- Minstrell, J. (2001). Facets of students' thinking: Designing to cross the gap from research to standards-based practice. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for science: Implications for professional instructional, and everyday science* (pp. 415–443). Mahwah: Erlbaum.
- Nathan, M. J., & Koedinger, K. R. (2000). An investigation of teachers' beliefs of students' algebra development. *Cognition and Instruction*, 18(2), 207–235.
- National Academy of Education. (2009). *Standards, assessments, and accountability: Education policy white paper*. (L. Shepard, J. Hannaway, & E. Baker, Eds.). Washington, DC: Author.
- Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
- Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19.
- Nilsson, N. (2007). The physical symbol system hypothesis: Status and prospects. In M. Lungarella (Ed.), *50 years of AI* (pp. 9–17). Berlin: Springer.
- Novick, L. R. (1990). Representational transfer in problem solving. *Psychological Science*, 1, 128–132.
- Ogan, A., Aleven, V., & Jones, C. (2010). Advancing development of intercultural competence through supporting predictions in narrative video. *International Journal of Artificial Intelligence in Education*, 19(3), 267–288.
- Ohlsson, S., & Mitrovic, A. (2007). Fidelity and efficiency of knowledge representations for intelligent tutoring systems. *Technology, Instruction, Cognition and Learning*, 5(2), 101–132.
- Papert, S. (2000). What's the big idea: Towards a pedagogy of idea power. *IBM Systems Journal*, 39, 3–4.
- Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). Organizing instruction and study to improve student learning (NCER 2007–2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.
- Pirolli, P. L., & Anderson, J. R. (1985). The acquisition of skill in the domain of programming recursion. *Canadian Journal of Psychology*, 39, 240–272.
- Polk, T. A., & Newell, A. (1995). Deduction as verbal reasoning. *Psychological Review*, 102(3), 533–566.
- Polya, G. (1957). *How to solve it: A new aspect of mathematical method*. (2nd ed.). Princeton, NJ: Princeton University Press.
- Post & Brennan (1976). An experimental study of the effectiveness of a formal versus an informal presentation of a general heuristic process on problem solving in tenth-grade geometry. *Journal for Research in Mathematics Education*, 7(1), 59–64.
- Ritter, S., Kulikowich, J., Lei, P., McGuire, C. L., & Morgan, P. (2007). What evidence matters? A randomized field trial of cognitive tutor algebra I. In T. Hirashima, U. Hoppe, & S. S. Young (Eds.), *Supporting learning flow through integrative technologies* (Vol. 162, pp. 13–20). Amsterdam, Netherlands: IOS Press.
- Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin and Review*, 2, 249–255.
- Rittle-Johnson, B., & Star, J. (2009). Compared to what? The effects of different comparisons on conceptual knowledge and procedural flexibility for equation solving. *Journal of Educational Psychology*, 101(3), 529–544.
- Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Learning and Verbal Behavior*, 6, 855–863.
- Renkl, A., & Atkinson, R. K. (2010). Learning from worked-out examples and problem solving. In J. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory and research in educational psychology* (pp. 91–108). New York: Cambridge University Press.
- Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1, 181–210.
- Roll, I., Aleven, V., & Koedinger, K. R. (2009). Helping students know 'further' - increasing the flexibility of students' knowledge using symbolic invention tasks. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference*

- of the cognitive science society* (pp. 1169–1174). Austin, TX: Cognitive Science Society.
- Roll, I., Aleven, V., & Koedinger, K. R. (2011). Outcomes and mechanisms of transfer in invention activities. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2824–2829). Austin, TX: Cognitive Science Society.
- Schoenfeld, A. H. (1985). *Mathematical problem solving*. Orlando, FA: Academic Press.
- Schoenfeld, A. H. (2007). Problem solving in the United States, 1970–2008: Research and theory, practice and politics. *ZDM Mathematics Education*, 39, 537–551.
- Schunn, C. D., & Anderson, J. R. (1999). The generality/specifity of expertise in scientific reasoning. *Cognitive Science*, 23, 337–370.
- Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, 22(2), 129–184.
- Schwartz, D. L., Martin, T., & Pfaffman, J. (2005). How mathematics propels the development of physical knowledge. *Journal of Cognition and Development*, 6(1), 65–88.
- Siegler, R. S. (2002). Microgenetic studies of self-explanation. In N. Granott & J. Parziale (Eds.), *Microdevelopment - transition processes in development and learning* (pp. 30–58). New York: Cambridge University Press.
- Singley, K., & Anderson, J. R. (1989). *The transfer of cognitive skill*. Cambridge, MA: Harvard University Press.
- Stigler, J. W. (1984). The effect of abacus training on Chinese children's mental calculation. *Cognitive Psychology*, 16, 145–176.
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12, 257–285.
- Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, 2, 59–89.
- Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review*, 8, 247–261.
- Tobias, S., & Duffy, T. M. (2009). *Constructivist theory applied to instruction: Success or failure?* New York: Routledge.
- Tomasello, M. (2000). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.
- VanLehn, K. (1987). Learning one subprocedure per lesson. *Artificial Intelligence*, 31, 1–40.
- VanLehn, K. (1999). Rule learning events in the acquisition of a complex skill: An evaluation of Cascade. *Journal of the Learning Sciences*, 8(1), 71–125.
- Van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, 17(1), 147–177.
- Yaron, D., Karabinos, M., Evans, K., Davenport, J., Cuadros J., & Greeno, J. (2010). Learning chemistry: What, when, and how? In M. K. Stein & L. Kucan (Eds.), *Instructional explanations in the disciplines* (pp. 41–50). New York: Springer Science+Business Media.
- Zhu, X., & Simon, H. A. (1987). Learning mathematics from examples and by doing. *Cognition and Instruction*, 4(3), 137–166.

INDEX

A

abduction, 149–150
causation and, 149–150
explanation and, 268
inconsistency resolution and, 150
similarity assessment and, 163, 164
abductive inferences, 268
model for, 269
reasoning and, 270–271
abductive reasoning
in medical reasoning, 739
in scientific thinking and reasoning, 707
absolute truth, in argumentation, 279
abstraction, in creativity, 468, 469
abstract selection task, 438
accounting. *See* behavioral accounting
accuracy, in outcome-motivated thinking, 398–399
circumstance distinction for, 399
cognitive-resource constraints on, 398
for nondirectional outcomes, 395, 405
reasoning and, 398–399
achievement, as genius, 493–494
acquire-a-company problem, 352
activation spreading, in formal thought disorder, 681–682
ACT-R computer model, 82
knowledge transfer with, 793–794
in modular components, 793–794
philosophy tenets for, 793
Adams v. New Jersey Steamboat Company, 727
adaptation, in analogical inference, 250
ad baculum argument, 282–283
additional intelligences, in general intelligence perspective, 497
additive counterfactuals, 401–402
ad hominem argument, 282
adjustment heuristic, 332. *See also* anchoring
Administrative Behavior (Simon), 756
ad populum argument, 282
ad verecundiam argument, 282

affect displays, 632
affect heuristic, 339. *See also* emotion, decision making and
aggregate field theory, 68
aging, thinking and
age-related cerebral atrophy and, 653
analogical reasoning and, 657–658
associative memory and, 660
cognitive theories for, 650–653
decision making and, 653–656
educational factors for, 657
episodic memory and, 660
expertise and, 663–664
explicit memory and, 660
false memories with, 660–661
framing effect and, 655
in general slowing theory, 651
IGT and, 654–655
induction and, 658–659
in inhibit deficit theory, 652
JOLs and, 662–663
judgment and, 653–656
memory and, 660–662
metacognition and, 662–663
metamemory and, 662–663
motivational factors for, 655
own-age bias and, 657
positive feedback and, 657
prefrontal theories for, 652
for problem solving, 656–657
reasoning and, 657–658
in reduced resources theory, 651
RPM test and, 657
semantic memory and, 660
set-shifting and, 659
Skinner on, 666
in SOC theory, 652–653
in SST, 653, 661
successful, 665–666
time assessment with, 655–656
training and, 664–665
with Wisconsin Card Sorting Task, 659
wisdom from, 665–666

working memory and, 660
Aha! reaction, 476
as unconscious, 479
AI. *See* artificial intelligence
algebra, as symbolic language, 798, 803, 803
algorithms
of analogical inference, 249
analogical reasoning models, 253–254
in analogy models, 240
explanations and, 265
in problem solving, 419
representations and, 37
Alighieri, Dante, 494
alignment-based models, for similarity assessment, 155, 165–167
alignable differences in, 167
feature matching in, 165
non-alignable differences in, 166–167
SIAM, 166
structural accounts of, 167
transformational models and, 168–169
allocentric representations, 614
alternating-offer bargaining, 353–355
anagrams, 482
analogical inference, 249–251
adaptation processes and, 250
algorithm of, 249
causal model integration with, 249–250
computational level analysis of, 253
as false memories, 250–251
function of, 249
representations and, 250
analogical reasoning
aging and, 657–658
in AI, 238, 238
algorithmic model development, 253–254
analogical inference, 249–251
analog retrieval in, 241
case-based reasoning as, 238
component processes of, 243–253

analogical (*cont*)

computational level analysis of, 253
definition of, 234
explicit processing in, 245
functions of, 235
implicit processing in, 245
by judges, 731
in legal reasoning, 727–728
with LISA, 241, 241
in long-term memory, 244–245
mapping in, 46, 235, 246
memory activation in, 245–246
neural implementation of, 254
processes of, 235
relational priming in, 245–246
relational structure in, 244–245
representations in, 253–254
retrieval gaps in, 244
schemas in, 251–253
in scientific thinking and reasoning, 709, 708–709
source analogs in, 243–244
sources in, 709
in structured relational representations, 45–46
targets in, 709
transfer paradigms in, 243–244
translational research for, 254–255
verbal, 242
working memory and, 240–241
analog problem solving, 421
analog mental representations, 39
analog presentations, in symbolic arithmetic, 599–600
analog representations, 39
mental, 39
analogy, 45–46
algorithmic models of, 240
applications for, 235
asymmetry in, 234
in cognitive psychology, 238
comparison in, 234
components of, 234
in consumer behavior research, 759
in creativity, 469
definition of, 234
empirical, 234
expert recognition of, 423
4-term, 236
as induction process, 235
knowledge representation and, 238–239
in mental model theory, 136
metaphor and, 236–238
multiconstraint theory and, 239
for number-line estimations, 598
parallel connectivity and, as constraint, 46
problem solving applications for, 238–239, 254
proportional, 236
in psychometric tradition, 236
relational reasoning and, 239–243

relationships in, 234

research history for, 236–239
in role-based relational reasoning, 235
in scientific thinking and reasoning, 709, 708–709
symbolic connectionism and, 240
systematicity principle and, 46
theories of, 46
analogy recognition, 423
analysis, computation level of, 4
analytic solving, 475, 478. *See also* problem solving
anchoring
numeric priming from, 333
paradigms for, 332
in positive model, for heuristics and biases approach to judgment, 328, 332–333
psychological processes from, 332–333
semantic activation from, 333
Anderson, John, 3, 53, 793
anger
decision making and, 314
as moral emotion, 369
animals
magnitude processing by, 589–592, 601
relational reinterpretation hypothesis for, 536, 537
in spatial orientation models, 558–559
teaching among, 533–534, 534
visuospatial thinking in, 606–607
ANT. *See* Attention Network Test
anti-representationalism
in cognitive psychology, 49
in embodied cognition, 48–49
in knowledge representation, 48–49
in situated cognition, 48–49
aphasia, 678
apologist position, 447
arationality, 434
argumentation
absolute truth in, 279
ad baculum, 282–283
ad hominem, 282
ad populum, 282
ad verecundiam, 282
applications of, 290–292
argument quality, 280–282
Bayes theorem for, 280, 292, 292
circular, 282
claims in, 278–279
for cognitive psychologists, 281
coherence in, 292
computer science applications for, 278
conditionals in, 286–290
content as factor in, 279–280
definition of, 277
in developmental psychology, 277–278
dialectical model of, 279
epistemic, 278, 282
with equivocation, 282
evaluation norms for, 291
external aspects of, 277
fallacies of, 282–290
function of, 277, 277
future research applications for, 292–293
from ignorance, 282, 283–284
inference in, 278
legal applications of, 279, 291–292, 293
logic in, 278
norms for, 278
in persuasion research, 277
philosophical foundations for, 278
pragma-dialectical theories of, 279
procedural aspects of, 281–282
psychology of, 280–292
SSA, 282, 288–289
theoretical frameworks of, 278–280
two-sided arguments, 281
Aristotelian explanations, 265
Aristotle, 236
rationality for, 434
arithmetic
for infants, 592–593
neuroimaging for, 592
non-symbolic numerical processing and, 594
symbolic, 599–600
artifacts, in mental model theory, 136
artificial categorization, 180
artificial intelligence (AI)
analogical reasoning in, 238, 238
case-based reasoning, 238
medical reasoning with, 737–738
normative reasoning theories and, 19
in transformational models, 167
associative approach, in causal learning, 213–215
augmented variants of, 213
Bayesian inferences compared to, 224–225
blocking in, 214–215
causal power theory and, 223
comparator hypothesis in, 215
contingency contrast in, 213–214
as intuitive, 213
with multiple causes, 214–215
with RW model, 214
sometimes-opponent-process model, 215
statistical model for, 213–214
associative memory, aging and, 660
asymmetric analogy, 234
asymmetric dominance, 307
asymmetric matching pennies game, 351
payoff predictions for, 351–349
asymmetric predictions, for similarity assessment, 162
asymmetric similarity, from contrast models, 162

- asymmetry predictions, for causality, 215–216
- atomistic concepts, 192–193
compositionality in, 192
- comprehension and interpretation in, 193
- attention
ANT for, 98
bounded, 3
cholinergic gene systems and, 100
cognitive neurogenetics for, 98–100
context, 682–683
in directional outcomes, 393–394
dopaminergic components of, 99–100, 98–99
glutamatergic genes and, 100
in inhibit deficit theory, 652
for insight problems, 484
for language of motion, 554
MSIT for, 105
for music, 784
neural architecture for, 98
in reduced resources theory, 651
selective, 682–683
serotonergic genes and, 100
visual, 339–340
to visual imagery, 611
working memory and, 98
- attentional defocusing, 466
- attentional processes, in moral judgment, 375
- Attention and Effort* (Kahneman), 339–340
- Attention Network Test (ANT), 98
- attributes, 43
directional outcome-motivated thinking influenced by, 392
in nondirectional outcomes, 395–396
- attribute evaluation, in decision making, 309–311
for compatibility, 309–310
evaluability effects in, 310
inconsistency in, 310–311
separate and comparative, 310–311
- attribution substitution account, 340, 340–341
in heuristics, 373
- auditory perception principles, 783
- auditory scene analysis, 776
- auditory sequence discrimination, 587
under Weber-Fechner law, 587
- auditory stream segregation, 776
- authority of law, for judges, 731–732, 731–732
- authority/respect domain, in moral judgment, 365
- autism, diagnosis of, 635
The Autonomous Set of Systems (TASS), 440
- availability heuristic, 330–331
- B**
- backward blocking, 215
- backward-driven reasoning. *See*
- hypothesis-driven medical reasoning
- BACON program, 500–501
- basic level categories, 180
- basic level categorization, 180
- Bayesian inferences, 23, 223–224
applications for other disciplines, 31–33
associative approach to, causal learning compared to, 224–225
- bias-variance tradeoff and, 27
- casual support models for, 223, 223–224
- in causal power theory, 223
- in cognitive modeling, 25
- computational levels of, 24–25
- future applications for, 33
- HBM, 29–30
- heuristics and, 25, 26
- inductive inferences and, 22–23, 29
- intervention in, 226–227
- JDM view for, 25–26, 26, 27
- levels of, 23–25
- mechanistic level of, 24–25
- methodological applications of, 223
- Monte Carlo methods for, 33
- normative models for, 25
- parsimony in, 224
- probability theory and, 16, 22, 23
- in rational models, 324
- rational process models for, 25
- in reasoning, 27–28
- reverse engineering and, 25, 26–27
- solution deviations for, 27
- statistics in, integration of, 28
- symbol integration in, 28
- in thinking, 27–28
- Bayes theorem, for argumentation, 280, 292, 292
- in arguments from ignorance, 284, 285
- in epistemic arguments, 284–285
- in medical reasoning, 737
- probability in, 293
- rationality and, 438
- BDNF gene, 100–101
- Beethoven, Ludwig van, 492
- behavioral accounting, 767–768
expertise in, 767–768
fact presentation in, 768
knowledge transfer prediction for, 768
numerical cognition for, 768
research goals for, 767
- Behavioral Decision Theory, 326
- behavioral finance, 765–767
decision making for, 765–767
expertise in, 767
goals of, 765
judgment for, 765–767
learning in, 767
in prediction markets, 766–767
for retirement, 765–766
in stock markets, 766
- vivid information for, 766
- behavioral game theory, 359
repeated play in, 360
- beliefs
culture and, 578–579
essentialism, 196–197
explanation and, 261
in featured representations, 42–43
logic and, 12, 18
probability and, 12, 16
sortalism, 197–198
- belief bias
in dual-process theory for deductive reasoning, 120, 123, 127–128
effect studies for, 81
- Bell, Victoria, 144
- Berliner, Hans, 493
- Bernoulli, Daniel, 302
- Berra, Yogi, 322
- bias. *See also* belief bias
cognitive, intelligence and, 122
confirmation, 577
in culture, as sampling bias, 573–574
in domain-general cognitive theories, 374
intention and, in moral judgments, 381
in judgments, 322–323
in MDS models, 161
in organizational behavior, 760
own-age, 657
in reasoning, 514
in representation of culture, 573
from status quo, 306
in transmission of culture, 573
in type 1 processes, 126
- bias-variance tradeoff, 27
- The Big Book of Concepts* (Murphy), 177
- big ideas, in knowledge transfer, 791–793, 798
vertical transfer of, 793
- binary predicates, 43
- birds-and-trains problem, 415
- bistable mapping, 247
- Bleuler, Eugen, 678–679
conceptualization of psychosis pathology, 679
- blind variation perspective, for genius, 501–504
advantages of, 502–503
creativity and, 501
cultural influences on, 502–503
domain expertise and, 503
empirical research on, lack of, 503–504
explanatory power of, 502
ideational variants in, 501
as integrative, 502, 503
limitations of, 503–504
predictive value of, 503
underdevelopment of, 502
- Blink* (Gladwell), 128

- blocking, 214–215
 backward, 215
 board games. *See* number board games
 Bohr, Niels, 322
 Bonaparte, Napoleon, 494
 Boroditsky, L., 561, 563
 bounded attention, 3
 bounded rationality, 324, 324, 327
 Bovens, Luc, 292
 Boyd, R., 533
 Boyle's law, 218–219
 Brahe, Tycho, 148, 707
 brain. *See also* prefrontal cortex, of brain
 domain specificity in, 199–200
 facial recognition domain in, 199–200
 under fMRI, 79–80
 frontal lobes, role in intelligence, 84–85
 hippocampal volume, schizophrenia and, 688–689
 insight in, regional activation for, 476
 insula, 355
 memory domains in, 200–201
 in monkeys, 530
 place and landscape recognition domain in, 200
 precuneus, 355
 temporal lobe abnormalities, schizophrenia and, 688–689
 Braine, Martin, 118
 brain size
 encephalization and, 530
 in humans, 530–531
 in monkeys, 530
 neocortex and, 531
 in primates, relative to other species, 530
 social brain hypothesis and, 531
 brainstorming, 465
 Brander, James, 351
 brand extension, 758
 Broca, Paul, 68
 Broca's area, 68
 Bromberger, Sylvain, 230
 Bruner, Jerome, 4, 703
 Buonarroti, Michelangelo, 492
 burden of proof, 279
 for juries, 729
 Burns, Robert, 494
 business, reasoning in
 in behavioral accounting, 767–768
 for behavioral finance, 765–767
 in business schools, 755
 future applications for, 769
 in management science, 763–765
 in marketing, 756–759
 in organizational behavior, 759–763
 business schools, 755
 future research goals for, 769
- C**
- Cage, Nicholas, 576
 Campbell, Donald, 501–502
- candle problem, 417
 capacity, in children's developmental thinking, 518–520
 abstract conceptualization, 519–520
 event representation, for infants, 518–519
 learning concepts in, 520
 methodology for, 518
 object representation in, 518–519
 capacity allocation, 683–684
 schizophrenia and, 684
 speech disorders and, 684
 Carey, Susan, 198
 Carlyle, Thomas, 494
 Carruthers, P., 704
 case-based reasoning, 238
 casual support models, in Bayesian inferences, 223, 223–224
 CAT. *See* computer axial tomography
 categorical induction, 707
 categorical reasoning, 619
 categorization
 artificial, 180
 basic level, 180
 causal learning and, 230
 characteristic comparisons within, 180
 cognitive neuroscience and, 187–189
 within concepts, 178
 connectionist models for, 185
 in consumer behavior, 757–759
 creativity and, 457–459
 decision bound models for, 184–185
 distortions within, 183–184
 exemplar views of, 184, 184
 explanation in, 261, 270
 feature learning from, 187
 features of, 187–189
 for goal planning, 178
 of graphical displays, 620
 implicit learning from, 186
 inference and, 183
 language function and, 189, 545
 learning from, 178, 179
 in legal reasoning, 725–726
 mixed models for, 185–186
 multiple function sensitivity with, 202
 multiple systems for, 185–186
 neuropsychological studies of, 187–188
 in number-line estimation, 598–599
 of numbers, 598–599
 of objects, in visuospatial thinking, 608
 of objects, language for, 551
 in organizational behavior, 761–762
 perceptual fluency and, 186
 perceptual-functional hypothesis and, 187–188, 188
 prototypes for, 183–184
 psychology of, in medical reasoning, 741–742
- rational models for, 184
 reasoning and, 189
 in semantics, 182–183
 sortalism and, 197
 of substances, language for, 551
 in surface-based problems, 423
 visual agnosia and, 187–188
 Cattell, Raymond, 236
 Cattell/Horn/Carroll (CHC) theory of intelligence, 443
 causal approach, in causal learning, 218–224
 acyclic constraints in, 218
 causal power theory for, 218–220
 ceiling effects in, 220–221
 experimental tests of, 221
 independent assumption effects in, 221–223
 normative models for, 221–222
 variants in, 218
 causal explanations, 265
 causal inference, 215–218
 Bayesian inferences and, 223–224
 causality in, as concept, 215–216
 ceiling effects in, 216–217
 diagnostic, 226–227
 experimental design principles with, 217
 inhibitory cues in, 216
 intervention in, 216, 217–218, 226–227
 invariances in, 217
 placebo effects with, 216
 causality, as concept, 215–216
 asymmetry predictions in, 215–216
 in children's developmental thinking, 521
 causal learning. *See also* associative approach, in causal learning; causal approach, in causal learning
 associations compared to, 211
 in associative world, 210–211
 assumptions in, 212
 Bayesian inferences and, 223–224
 category formation and, 230
 causality constraints in, 227–229
 cause and effect variables in, 211–212, 227–228, 228–229
 ceiling effects in, 220–221
 contiguous events in, 228
 definitions in, 211
 domain integration functions, 225–226
 domain-specific biases in, 212
 explanations for, 211
 function of, 210–211
 Hume and, 213
 hyperbolic discounting in, 228
 hypothesis revision for, 230
 independent causal assumptions in, 225–226
 intervention in, 213

- judgments from, 230
 in non-associative world, 210
 parsimony in, 230
 psychological theories for, 227
 time constraints in, 227–229
 time frame variability in, 228
 universality of explanations from, 211
 causal maps, for inductive inferences, 28
 causal model theory, for moral judgment, 374–375
 causal power theory, 218–220
 alternative causes in, 219
 associative approach in causal learning and, 223
 Bayesian inferences in, 223
 belief assumptions in, 219–220
 Boyle's law and, 218–219
 confounding conditions in, 220
 multiple causes in, 219
 nonconfounding conditions in, 220
 processing capacity for, 219
 causal thinking, 706–707
 mechanism in, 706
 in medical reasoning, 744–746
 for unexpected findings, 706–707
 ceiling effects
 in causal inference, 216–217
 in causal learning, 220–221
 CHC theory of intelligence. *See* Cattell/Horn/Carroll theory of intelligence
 children. *See also* capacity, in children's
 developmental thinking;
 developmental thinking in children
 egocentrism in, 515, 516
 as geniuses, 495
 gesture for, 633
 language development for, gesture and, 633–634, 634–635
 mathematical equivalence by, through gesture, 636
 physics for, 711
 scientific thinking and reasoning in, 710–712
 similarity assessment among, 171
 in SMPY, 505
 task mastery by, gesture and, 635–636
 theory change for, 521
 CH model. *See* cognitive hierarchy model
 cholinergic gene systems, attention and, 100
 Chomsky, Noam, 532
 universalist perspective on language, 544
 chords, 775
 chromatin, 94
 chromosomes, schizophrenia and, 688
 circular argument, 282
 as conditional, 289–290
 dependency in, 289
 equivalency in, 289
 probability theory and, 290
- circumstance distinction, in outcome-motivated thinking, 399
 civil law nations, 733
 cladograms, 426
 claims, in argumentation, 278–279
 clinical knowledge, in medical reasoning, 742, 745
 closure motivations
 for nondirectional outcomes, 397
 in outcome-motivated thinking, 399
 CNV. *See* contingent negative variation
 co-construction of knowledge, for
 culture, 573
 in history, 573
 reproduction of, 573
 code-switching, 561
 cognition. *See also* creativity, cognition
 and; implicit cognition, creativity and; process models, of higher cognition
 cultural cognition, evolution of, 576–577
 for culture, 576–579
 through culture, 571–576
 culture and, 569–570
 embodied, 47–48
 explanation and, 261
 extended mind hypothesis and, 575
 group, 763
 language and, 531–532, 543
 music as influence on, 779
 neurogenetics for, 94–98
 situated, 48
 structural mapping theory for, 465
 working memory and, 95
 cognition as proof theory, 15
 cognitive-affective theory, 368
Cognitive Basis of Science (Carruthers/Stich/Siegal), 704
 cognitive bias, intelligence and, 122
 type 1 processes and, in dual process theory, 126
 cognitive control, in schizophrenia, 681
 cognitive deficits, in formal thought disorder, 684
 cognitive development
 ACT-R computer model for, 82
 explanation in, 267, 268
 gesture and, 633
 inductive inferences and, 29
 theoretical history of, 3
 cognitive division of labor, 576
 cognitive economy, 179
 cognitive hierarchy (CH) model,
 347–348, 348–358
 advantages for, 348
 applications for, 348
 for asymmetric matching pennies game, 351
 for entry games, 351–352
 eye-tracking evidence for, 353–355
 field data for, 356–358
 fMRI imaging for, 355–356
- individual differences in, 355–356
 individual game structures and, 355–356
 payoff predictions in, 350–351
 for PBC games, 348–351
 for private information games, 352–353
 cognitive illusion paradigm, 327
 in positive model, for heuristics and biases approach to judgment, 329
 cognitive illusions, 468
 cognitive impairment. *See* thought disorders
 cognitive load theory, 797
 in data-driven reasoning, 741
 cognitive maps, 614
 cognitive miser model, for cognition, 337–338
Cognitive Models of Science (Giere), 704
 cognitive neurogenetics
 for attention, 98–100
 in cognitive neuroscience, 92–103
 in complex cognition, 94–98
 development of, 90
 DNA and, 94
 epigenetic influences on, 94
 future applications for, 103, 104
 intermediate phenotype approach in, 92
 interpretive issues in, 92–94
 for language cognition, 102–103
 for long-term memory, 100–102
 methodology issues with, 93–94
 molecular biogenetic data integration in, 104
 molecular-biological sequence in, 91–92
 neuroimaging and, 104
 new data for, 103–104
 nucleotides in, 91–92
 path-analytic approach in, 105
 phenotypes in, 92
 polygenicity in, 92
 psychological theory and, 104
 sample size guidelines for, 104
 systems approach in, 105
 in working memory, 94–98
 cognitive neuropsychology, 68–73
 aggregate field theory in, 68
 behavioral study in, 68–72
 with brain-damaged patients, 72–73
 localizationism theory in, 68
 cognitive neuroscience, 3
 academic development of, 67–68
 categorization and, 187–189
 cognitive neuropsychology, 68–73
 creativity and, research applications of, 470–471
 experimental methods for, 69–72
 mental model theory and, 136
 for multiple memory systems, 185
 neurogenetic research in, 92–103

- cognitive neuroscience (*cont.*)
 with neuroimaging, 73–81
 similarity assessment in, 156–157,
 170–171, 171–172
 terminology for, 86
- cognitive psychology
 analogy in, 238
 anti-representationalism in, 49
 argumentation in, 281
 knowledge representation in, 47
- cognitive science
 children's developmental thinking
 and, 524
 Great Rationality Debate in, 433,
 435–436
 irrationality and, 434
- cognitive support systems, in management
 science, 763–764
- Cognitive Tutors systems, 793
- Cohen, Jonathan, 325, 679
 psychological probability study, 329
- coherence
 in analogical mapping, 247
 in argumentation, 292
 multiconstraint theory for, 247
 in rational models, 323
- Coke, Edward, 719
- collaborative cognition, 470
- collaborative memory, 571–572
- Collins, Allan, 181
- color perception, in language, 549–550
 cross-linguistic commonalities for, 550
 labeling of, 549–550, 563
- common law nations, 733
- communication, concepts and, 178
- communication deviance, 690–691
- comparative attribute evaluation, in
 decision making, 310–311
- comparator hypothesis, 215
- compatibility, in decision making,
 309–310
 preferences for, 309
 response modes and, 309–310
- competence/performance distinction, for
 rationality, 448
- compiled causal knowledge, 746
- complete explanation, 261–262
 selected compared, 262
- compositionality, 192
- compound premises, in deductive
 reasoning, 143
- compound remote associate (CRA)
 problems, 476
- compromise effect, 307–308
- computation
 analysis and, 4
 function of, 4
- computational models. *See also* process
 models, of higher cognition
 ACT-R, 82
 for concepts, future applications of,
 202
 for knowledge transfer, 793–794
- LISA, 61–62, 82–84
 of neural systems, 82–84, 85
 for symbolic languages, 795
- computer axial tomography (CAT),
 73–74
- computers, for scientific thinking and
 reasoning, 712
- computer science, argumentation
 applications in, 278
- computer simulations, for human
 thinking, 3
- concepts
 abstract, in children's developmental
 thinking, 519–520
 atomic, 192–193
 brain region domains for, 199–201
 categorization within, 178
 causality, 215–216
 cognitive divisions between, 195–196
 combination of, 178, 190–192
 communication and, 178
 comprehension of, 178
 creativity for, 457–459
 definition of, 177
 development of, as process, 180–181
 domain specificity for, 198
 early research history on, 179
 essentialism, 196–197
 evidentiality, language, 555–556
 facial recognition, 199–200
 family resemblance between, 180, 180
 in folk biology, 198–199
 formation of, 457–458
 functions of, 178–179
 future applications, through
 computational models, 202
 generic noun phrases, 193–194
 gesture and, demand for, 639
 hierarchical structure of, 458
 inferences from, 458–459
 inferential, 192–193
 interdomain differences for, 199
 knowledge elements in, 710
 language function and, 189
 in long-term memory, storage of, 201
 memory organization and, 178–179
 in memory structures, 189–190
 modality specificity for, 188
 motion, language for, 553–554
 neuroimaging studies for, 188–189
 number, language for, 557–558
 object, 188
 organization of, 458
 participant population diversity, for
 research applications, 202
 path-of-least-resistance theory and,
 459
- polysemy, 194–195
- prediction from, 178
- psycholinguistic research and, 202
- psychological study of, 177
- in psychometaphysics, 178, 195, 202
- in riskless decision making, 306
- in scientific thinking and reasoning,
 development of, 709–710
 semantic memory and, theories for,
 179–182
- sortalism, 197–198
- spatial frames of reference as, language
 for, 554–555
- spatial orientation as, language for,
 558–561
- spatial relationships as, language for,
 552–553
- substances, language for, 550–552
- in thinking, 1
- time, language for, 556–557
- concept combinations, 178, 190–192
 distinctions within, 192
 predictions for, 192
 selective modification model for, 193
 theory development for,
 191–192, 191
 typicality of structures in, 190–192
- conceptual combination, 178, 190–192
 in creativity, 464–465, 469
 emergent properties in, 465
- conceptual knowledge, through gesture,
 633
- conceptual processing, in embodied
 cognition, 48
- conditional inference, 118, 122–123
 in mental model theory, 119
- conditionals
 in argumentation, 286–290
 circular argument as, 289–290
 deontic, 288
 in fallacies, 286–290
 SSA as, 288–289
 utility, 288
- confirmation bias
 culture and, 577
 in hypotheses testing, 705
 legal reasoning and, 720–721
 variants for, 705–706
- conflicts
 decisional, 307–308
 in decision making, 306–309
 reason-based choice and, 308–309
- conformity, in creativity, 468
- confounding conditions, in causal power
 theory, 220
- congruency effect, 608
- conjunction fallacy, for judgment
 heuristics, 337
- probability theory and, 337
- conjunction rule, 337
- conjunctive coding, 58
 in vector addition symbolic-
 connectionist models, 62
- connectionist models
 for categories, 185
 meaning-based memory in, 201–202
 symbolic, 60–64
 vector addition, 61–63
 vector multiplication, 60–61

- connectionist neural-network models, 55–60
- auto-associative networks in, 57
- conjunctive coding in, 58
- connections in, 56–57
- degradation within, 57
- degrees of freedom in, 58
- disadvantages of, 58–60
- distributed representations in, 56, 56
- feed-forward networks in, 57
- future applications of, 64
- individual mapping in, 58
- input/output units in, 59
- integration within, 57
- localist representations in, 56
- mathematical simplicity of, 57
- processes in, 56–57
- query units in, 59
- recurrent networks in, 57
- relational processing in, 59
- representations in, 56
- strengths of, 57
- Connery, Sean, 576
- consciousness
- private body, 368
 - thinking and, 2
- consequence, from logic, 14
- consistency
- in logic, 12, 14–15, 18
 - model-theoretic, 14
 - proof-theoretic, 14
 - as static, 15
- consistent equations, 425
- constructivist science education, 713
- consumer behavior, marketing and, 756–759
- analogy research for, 759
 - brand extension in, 758
 - categories in, 757–759
 - context for, 757
 - creativity and, 759
 - decision making in, 756–757
 - expertise in, 757
 - learning in, 757
 - preferences in, 758–759
 - research on, 758
 - similarity research for, 759
- contemporary recognition, posthumous *vs.*, 494
- context attention, in formal thought disorder, 682–683
- contingent negative variation (CNV), 76
- contradictions, in logic, 13, 13
- contrast models, in similarity assessment, 161–163
- asymmetric similarity predictions with, 162
 - common features in, 162, 162–163
 - computational development of, 163
 - premise of, 163
- conventional morality, 366
- convergence schemas, 251–252
- convergent thinking, 465
- cooperation, in organizational behavior, 762
- correlational psychology, genius study within, 505
- counterfactual effects, on strategy-motivated thinking, 401–402
- counting systems, cross-linguistic differences in, 557–558, 563
- cover and differentiate model, in medical reasoning, 739
- CQ test. *See* creativity quotient test
- Craik, Kenneth, 3, 135
- CRA problems. *See* compound remote associate problems
- creativity, cognition and. *See also* metrics of creativity; music; music composition
- abstraction in, 468, 469
 - aids to, 469–470
 - analogy in, 469
 - in blind variation perspective, for genius, 501
 - categorization and, 457–459
 - cognitive illusions and, 468
 - cognitive neuroscience applications, 470
 - collaborative cognition and, 470
 - conceptual combination in, 464–465, 469
 - conceptualization and, 457–459
 - consumer behavior and, 759
 - CQ test for, 495–496
 - creative cognition approach to, 456
 - Darwinian model of, 466
 - digital tools for, 470–471
 - domains of, 456, 457–464
 - expertise from, 471
 - false memories and, 457
 - as genius, 494
 - genius and, 494
 - idea roadmaps theory for, 467
 - ideation in, 465
 - impediments to, 468–470
 - implicit assumptions and, 468
 - implicit cognition and, 463, 464
 - implicit memory and, 459, 464
 - inadequate knowledge and, 468
 - influential factors for, 456
 - insight and, 487–488
 - intuition in, 464
 - knowledge storage in, 457, 469–470
 - in marketing, 759
 - measurement tests for, 495–496
 - memory and, 459–462
 - noticing in, 469
 - opportunistic assimilation theory of, 466
 - in organizational behavior, 762–763
 - path-of-least-resistance theory and, 459
 - premature conceptualization and, 469
 - prepared mind theory for, 466
 - problem solving and, 461, 462–463, 471
- remote association theory for, 465–466
- sketching and, 470
- structured imagination theory for, 457, 465
- support technologies for, 470–471
- as unexpected, 457
- wisdom and, 665, 665
- creativity quotient (CQ) test, 495–496
- Crick, Francis, 701
- Cromwell, Oliver, 494
- Cross, Ian, 774
- cross-linguistic differences, in language
- for counting systems, 557–558, 563
 - for evidentiality, 556
 - through gesture, 641
 - for motion, 553–554, 554, 563
 - for number, 557–558, 563
 - for spatial frames of reference, 554–555
 - for spatial relationships, 552–553
- cross-mapping, in analogical reasoning, 247
- cross-modal mapping, by infants, 588–589
- preferences in, 588
- crystallized intelligence, 84, 443
- mindware and, 446
- cultural cognition, 576–577
- mind-reading and, 576
 - neural systems for, 576–577
- cultural distribution, across peoples, 571–572
- collaborative memory and, 571–572
 - with transactive memory, 571
- cultural intelligence hypothesis, nonhuman primates and, 534–535
- culture
- belief systems and, 578–579
 - blind variation perspective on genius and, 502–503
 - co-construction of knowledge for, 573
 - cognition and, 569–570
 - cognition for, 576–579
 - cognition through, 571–576
 - collaborative memory and, 571–572
 - confirmation bias and, 577
 - cultural cognition, evolution of, 576–577
 - definition of, 570
 - developmental thinking in children and, 517
 - distributed across peoples, 571–572
 - dual-process theories influenced by, 124
 - dynamic social impact process in, 573
 - group polarization in, 574
 - human mind influenced by, 533–534
 - imitation as learning tool in, 572
 - information cascades in, 573–574
 - isolation costs, 574–575
 - knowledge in representation of, 572

culture (*cont.*)

- knowledge transmission of, 572
- through language, 543
- macropsychology and, 579–580
- moral grammar theory and, 372
- moral judgment influenced by, 365, 385–386
- motivated reasoning and, 577
- neural systems for cognition of, 576–577
- organizational behavior and, 762, 762
- pluralistic ignorance in, 574
- prediction games in, 574
- predispositions to, 577–578
- in rationalist theory, 367
- relational models theory and, 577
- relationship regulation theory and, 577
- representation and, across time, 572–573
- research models, 579
- sacred and protected values and, 382
- sampling biases for, 573–574
- SNPs and, susceptibility of, 580
- social-functional aspects of, 577–578
- technology and, 575–576
- thinking and, 570
- transactive memory and, 571
- transmission of, across time, 572–573

Curie, Marie, 500

D

- Darwinian model of creativity, 466
- data-driven medical reasoning, 740–741
 - cognitive load in, 741
- data mining, 712
 - for knowledge transfer, 799–800
- DCM. *See* dynamic causal modeling
- DDE. *See* doctrine of double effect
- decisional conflicts, 307–308
 - asymmetric dominance in, 307
 - compromise effect in, 307–308
 - default alternatives in, 307
 - among experts, 307
 - influences on, 308
 - status quo from, 307
- decision analysis, in medical reasoning, 737
- decision bound models, for categories, 184–185
- decision making. *See also* moral judgment
 - aging and, 653–656
 - anger and, 314
 - for behavioral finance, 765–767
 - compatibility in, 309–310
 - components of, 747
 - conflicts in, 306–309
 - in consumer behavior, 756–757
 - decision maker competence, 303
 - definition of, 301
 - descriptive models for, 316
 - disgust and, 314
 - domain specificity for, 200

drive and, 314–315

- emotion and, 314–315
- empathy gaps in, 315
- explanations in, 270
- frame of mind and, 313–315
- framing effect and, with aging, 655
- functions of, 301
- for future events, 313
- future research applications for, 316
- global perspectives for, 311–315
- identity and, 314
- IGT and, aging as influence on, 654–655
- influent factors for, 301
- in intensive care units, 747
- by juries, psychological model for, 730
- local perspectives for, 311–315
- in marketing, 756–757
- in medical reasoning, 746–747, 747–748
- mental accounting in, 312
- metacognitive influences on, 315
- monetary outcomes and, 303
- motivational factors for, with aging, 655
- NDM, 747
- nonmonetary outcomes and, 303
- normative theories and, as standard, 11–12
- for organizational behavior, 760
- preferences in, 316
- properties of, 2
- rational theory of choice and, 12, 16–18, 18, 301–302
- reason-based choice in, 308–309
- riskless, 305–306
- sadness and, 314
- under uncertainty, 302–305
- weighting of attributes in, 309–311
- decision support systems, in
 - management science, 764
- dedicated neural systems, 79
- deduction
 - inferences in, 2
 - in medical reasoning, 738
 - in mental model theory, 136–140
 - in thinking, 2
- deductive inferences, explanations and, 271
- deductive-nomological (DN) explanation, 262
- deductive reasoning
 - compound premises in, 143
 - conditionals in, 144
 - dedicated neural systems for, 79
 - dual-process theories for, 115, 119–120, 117–121
 - error prediction in, 141
 - experimental studies of, 140–144
 - fallacies in, 117, 124–129, 117–130
 - fMRI and, 78–79
 - illusory inferences in, 142–143, 143
 - inconsistencies in, 147–148
 - invalid inferences, 141–142
- mental logic and, 118
- mental model theory for, 118–119, 134
- with multiple models, 140–141
- by naïve individuals, 143–144
- PPA and, 78–79
- predictions for, 140–144
- psychology of, 115–116
- with quantifiers, 144
- relational, 144
- in scientific thinking and reasoning, 707–708
- suppositions in, 143–144
- visuospatial thinking in, 618–620

Dehaene-Changeux model, 590

- deictic gestures, 639
- delayed feedback, 764–765
- deliberation instructions, for juries, 729
- delusions, in schizophrenia, 675
- dementia, 684–685
- Dennett, Dan, 441
- density estimation, 31
- deontic conditionals, in fallacies, 288
- deontic logics, 15
- deoxyribonucleic acid (DNA)
 - chromatin, 94
 - as epigenetic influence, in cognitive neurogenetics, 94
 - histones, 94
- dependency circularity, 289
- Descartes, René, 1, 492, 492
- developmental psychology. *See*
 - also* capacity, in children's developmental thinking; children; infants; learning from others, among children
- analogy mapping in, 248–249
- argumentation in, 277–278
- gesture and, 633
- inductive inferences in, 32–33
- knowledge representation in, 47
- numerical symbol comprehension, 594–601, 601–602
- rationalist theory in, 366
- similarity assessment in, 169–170
- sortalism in, 197
- spatial orientation in, language for, 560–561
- spatial relationships in, language for, 563
- developmental thinking, in children
 - causality in, 521
 - causal reasoning in, 520–521
 - change stages in, 514–518
 - cognitive science applications for, 524
 - cultural variation in, 517
 - domain general principles for, 516–517
- domain specificity in, 516–517
- early capacity in, 518–520
- egocentrism and, 515, 516
- experience effects on, 517–518
- expertise and, 516

- incremental changes in, 515–516
 individual variation in, 517
 knowledge evolution from, 513
 language development in, 513–514
 in learning from others, 521–524
 as methodological tool, 514
 modularity in, 516
 for number, 557
 ontological commitment in, 520–521
 perceptual narrowing in, 517–518
 qualitative changes in, 515–516
 reasoning biases and, 514
 research goals for, 513–514
 reverse patterns for, 516
 skills assessment from, 513
 task analysis for, 515–516
 theory change in, 521
 training effects on, 517–518
 underestimation of abilities for, 515
- Devi, Shakuntala, 493, 498
- diagnostic causal inference, 226–227
- diagnosticity effect, 170
- diagrams, in problem solving, 425, 426–427
- cladograms, 426
- in perceptual principle, 426
- dialectical model, of argumentation, 279
- Dickinson, Emily, 494
- diffusion tensor imaging (DTI), 74
- dilemmas. *See also* moral dilemmas
- Heinz, 366
 - high-conflict, 370
 - trolley, 377–379
- directional outcomes, in motivated thinking, 391, 391–395
- attentional effects on, 393–394
- attribution effects on, 392
- information search effects on, 393–394
- knowledge activation effects on, 394–395
- legitimacy effects on, 393
- memory effects on, 395
- organization effects on, 395
- recall effects on, 394–395
- truth evaluation effects on, 393
- discourse, in mental model theory, 136
- Discourse on Method* (Descartes), 492
- discovery, intervention in, 213
- discovery learning approach, in science education, 713
- Discovery of Relations by Analogy (DORA), 62–63
- disgust, decision making and, 314
- dishabituation paradigm, 548
- disorientation, 616
- cues after, 616
- distributed representations, in
- connectionist neural-network models, 56, 56
- divergent thinking, 465
- DMPFC. *See* dorsomedial prefrontal cortex
- DNA. *See* deoxyribonucleic acid
- DN explanation. *See* deductive-nomological explanation
- doctrine of double effect (DDE), 372
- causal processes in, 379
- intention and, 380
- trolley dilemma and, 378, 379
- Doherty, Michael, 704, 706
- domain expertise, for genius, 498–499
- advantages of, 498
- in blind variation perspective, 503
- development of, 498
- empirical research for, 498
- limitations of, 498–499
- nature *vs.* nurture debate for, 498–499
- prodigious performance and, 498
- domain-general cognitive theories, for
- moral judgment, 374–375, 384
 - causal model theory, 374–375
 - contemporary research on, 385
 - excessive mental stimulation and, 375
 - fluency processing in, 375
 - framing effects in, 374
 - JDM, 374
 - motivational systems, 375
 - outcome bias in, 374
- domain specificity, for concepts, 198
- in brain regions, 199–200
 - for decision making, 200
 - in essentialism, 198
 - for general-purpose decision making, 200
 - in learning, 198
 - for memory, 200–201
- dominance solvable game, 351
- “do no harm” heuristic, 373–374
- dopamine-system-related genes
- in attention, 99–100, 98–99
 - in long-term memory, 101
- DORA. *See* Discovery of Relations by Analogy
- dorsolateral prefrontal cortex (DLPFC), 355
- dorsomedial prefrontal cortex (DMPFC), 355
- double causal contrast theory, 379–380
- drive, decision making and, 314–315
- DTI. *See* diffusion tensor imaging
- dual-process theories, for deductive reasoning, 117–121. *See also* type 1 processes, in dual process theory; type 2 processes, in dual process theory
- automatic/shallow information processing in, 115, 116
- belief bias in, 120, 123, 127–128
- cognitive architecture within, 125
- conditional inference in, 118, 122–123
- conscious processing in, 116
- contemporary research on, 129
- cross-cultural influences on, 124
- development of, 115, 119–120
- distinctions within, 124–125
- evidence sources for, 120
- fallacies in, 117, 124–129, 117–130
- future applications for, 129–130
- in heuristics and biases approach to judgment, 328
- hypothesis testing for, 118
- in judgment, 339–341
- origins of, 117–121
- ratio bias in, 340
- received view of, 121, 116–117
- selection tasks and, 119
- social knowledge in, 115
- specificity of proposals in, 125
- two systems in, 117, 130
- type 1 processes, 121–124
- type 2 processes, 121–124
- working memory and, 121–122
- dual-process theories, for moral judgment
- consequentialist responses with, 370
 - critiques of, 370–371
 - deontological responses with, 370
 - evidence for, 370
 - for high-conflict dilemmas, 370
 - versions of, 371
- dual-process theories, for rationality, 438–441
- secondary presentations in, 441
 - TASS, 440
 - type 1 processing in, 439–441
 - type 2 processing in, 439–441
 - ubiquity of, 438–439
- Duncker, Karl, 3, 413, 416. *See also* Gestalt psychology
- dynamic attending, 776
- dynamic causal modeling (DCM), 80
- dynamic social impact process, 573

E

- ECOG. *See* electrocorticography
- Edelman, Shimon, 158
- education. *See* business schools; constructivist science education; science education
- educational data mining, for knowledge transfer, 799–800
- Edwards, Ward, 325
- EEG. *See* electroencephalography
- egocentric representations, 614, 615
- egocentrism, in children, 515, 516
- Eimas, Peter, 548
- Einstein, Albert, 497, 703
- electrocorticography (ECOG), 75
- electroencephalography (EEG), 73, 75
- gesture under, 637–638
 - insight activation under, 476–478
 - mental number line under, 591–592
- electrophysiological functional neuroimaging, 74–77
- ECOG, 75
- EEG, 73, 75
- EROs, 76–77
- ERPs, 75–76

- electrophysiological (*cont.*)
 single and multiunit recording, 74–75
- Eliot, T.S., 497
- emblems, 632
- embodied cognition, 47–48
 anti-representationalism in, 48–49
 conceptual processing in, 48
 for vision, 47
- emergence, in creativity, 468
- emotions
 moral, 369, 369–370
 in PG framework, 358–359
 in response to music, 778–779, 786
- emotion, decision making and, 314–315
 affect heuristic for, 315
 anger, 314, 369
 disgust, 314
 empathy gaps from, 315
 guilt, 369–370
 inconsistency from, 315
 insight and, 484
 mood maintenance, 314
 for moral dilemmas, 380
 moral judgment and, 368–370, 384
 motivated thinking and, 404–405
 neuroimaging for, 369
 neuropsychological studies for, 369
 on prefrontal cortex, 369
 sadness, 314
 sentimental rules theory and, 369
 in SST, 653
 thin and thick moral concepts, 369
- emotion-focused problem solving, aging and, 657
- empathy gaps, in decision making, 315
- empirical analogies, 234
- empiricism, 30
- empiricist perspective, on language, 544
- emulation, imitation compared to, 522
- encephalization, brain size and, 530
 social complexity and, 531
- encoding, in language, 547–548
- endophenotype, for schizophrenia, 685, 685, 689–690
- endowment effects, 305, 306
- English common law, 723–724
- entry games, 351–352
 coordinating issues in, 351
 Nash equilibrium in, 352
- environmental space, 607, 613–618.
See also orientation, in
 environmental space
- allocentric representations in, 614
- cognitive processes for, 613
- egocentric representations in, 614, 615
- navigation through, 613
- perspective taking for, 615
- spatial layout knowledge and, 616–617
- spatial representations in, 614–615
- vision and, 613
- wayfinding through, 617–618
- episodic memory
 aging and, 660
 genetic polymorphisms in, 98
 retrieval of, in prefrontal cortex, 85
 schizophrenia and, 688–689
- epistemic argumentation, 278, 282
 Bayes theorem in, 284–285
 logic in, 282
- epistemic rationality, 434–435
- equations, in problem solving, 425–426
 consistent, 425
 fluency in, 426
 grouping of, 426
 inconsistent, 425
 perceptual learning modules and, 425–426
- equilibrium points, 347
- equiprobability principle, 144–145
 for naïve reasoners, 144
- equivocality circularity, 289
- equivocation argument, 282
- EROS. *See* event-related optical signal
- EROs. *See* event-related oscillations
- ERPs. *See* event-related potential errors
 classification of, 746
 fundamental attribution, 733
 in goals, 746–747
 heuristics as source of, 746–747
 of intention, 746–747
 in medical reasoning, 746–747
 prediction of, in deductive reasoning, 141
- essentialism, 196–197
 causal structures for, 197
 classifications within, 196
 domain specificity in, 198
 foundational philosophy of, 196
 indirect evidence for, 196
 induction potential of, 196
 sortalism and, 197–198
- ethics, in organizational behavior, 760
- EU. *See* expected utility
- evaluability effects, 310
- event-related optical signal (EROS), 81
- event-related oscillations (EROs), 76–77
 applications for, 76–77
 in time-frequency analyses, 77
- event-related potential (ERPs), 75–76
 applications for, 76
 CNV and, 76
 insight and, 480
- event representation, for infants, 518–519
- evidentiality, language for, 555–556
 cross-linguistic differences in, 556
- evolution, 436–437
 natural selection and, 436
 of rationality, among humans, 436, 436–437, 437
- evolutionary psychology, subset principle in, 145
- exaggerated imitation, 522
- excessive yielding to normal biases, in formal thought disorder, 683
- expected utility (EU)
 in prospect theory, 302–303
 in rational theory of choice, 17, 301–302
 SEU, 323
- expertise, problem solving and, 421–423
 aging and, 663–664
 analogy recognition and, 423
 in behavioral accounting, 767–768
 in behavioral finance, 767
 children's developmental thinking and, 516
 in consumer behavior, 757
 from creativity, 471
 in marketing, 757
 pattern recognition and, 422
 problem structure knowledge and, 422
 in SOC theory, 664
 stored solution plans and, 422
- explanations
 abduction and, 268
 abductive inferences and, 268
 adaptive behavior support from, 263–264
 in AI, 260
 at algorithmic level, 265
 Aristotelian, 265
 in categorization, 261, 270
 causal, 265
 for causal learning, 211
 causal mechanisms for, 263, 267
 causal theories for, 262–263, 263
 classifications for, 265
 cognition and, 261
 in cognitive development, 267, 268
 complete, 261–262
 at computational level, 265
 contrast in, 262
 as control mechanism, 263
 in decision making, 270
 deductive inferences and, 271
 definition of, 261–262
 DN account of, 262
 empirical support for, within psychology, 263
 events in, 267
- feature support for, 270
- functions of, 263–264
- future applications for, 271–272
- future learning and, 263
- generation of, 266–267, 272
- as heterogeneous, 271
- at implementational level, 265
- inductive inference and, 271
- inference and, 268–270
- during interactive cognitive activity, 266
- judgments and, 270
- language and, separation from, 272
- in learning, 265–268
- mechanistic, 264

- as phenomenologically satisfying, 263
 prior beliefs and, 261
 probability and, 268–269
 as process, 261, 262
 as product, 261, 262
 properties in, 267
 purpose of, 260
 questions and, 266
 reasoning and, 261, 270–271
 for recipients, 266
 research theory on, 261
 seeking of, 266
 selected, 262
 self-assessment of, 270
 self-explanation, 267, 268
 structure of, 262–263
 subsumption and, 267–268
 taxonomies in, 264
 teleological, 264
 types of, 264–265
 unification from, 267–268
 explanatory coherence model, for juries, 730
 explicit inductions, 146
 explicit memory, aging and, 660
 explicit reasoning
 processing speed for, 128–129
 selection tasks for, 120
Exploring Science (Klahr), 704
 extended mind hypothesis, 575
 extensional reasoning, 144
 extension effect, 170
 external representations, 425
 extreme Panglossianism, 447. *See also*
 apologist position
 eye-tracking measures, for CH game model, 353–355
 alternating-offer bargaining and, 353–355
 Mouselab, 353, 360
 eye-tracking technology, for fixations, 484
- F**
- facial recognition, brain region for, 199–200
 facts
 in behavioral accounting, 768
 for juries, 721–722
 law and, distinction between, 721–722
 psychology of determination for, 722
 fact finding, legal reasoning and, 728–731
 by judges, 730–731
 by juries, 729–730
 fairness/reciprocity domain, in moral judgment, 365
 fallacies
 of argumentation, 282–290
 conditional, 286–290
 in deductive reasoning, 117, 124–129, 117–130
 planning, in judgment heuristics, 336
 subjective degree of belief in, 286–287
- false memories, 457
 with aging, 660–661
 false memories, analogical inferences as, 250–251
 recognition paradigm for, 250
 source analogs in, 251
 Faraday's Law, 80
 featural models, for similarity assessment, 155, 160–161
 abduction reasoning and, 163, 164
 bias in, 161
 geometric models and, 163–165
 hierarchical cluster analysis in, 163–164
 hierarchical representations in, 164
 minimality assumption in, 160
 propositions in, 164
 symmetry assumption in, 160
 triangle inequality assumption in, 160–161
 unstructured representations in, 164
 featured representations, 42–43
 degrees of belief in, 42–43
 elements of, 42
 information in, augmentation of, 42
 for similarity studies, 42
 for speech perception, 42
- feature learning
 in categorization, 187
 similarity and, 187
 Fermat's Last Theorem, 13
 fetal hypoxia, 685
 figural space, 607
 finance. *See* behavioral finance
 fixation
 eye-tracking technology for, 484
 in idea roadmaps theory, 467
 insight and, 480
 in insight problems, 483–484
 Fleming, Renée, 780
 floating intentionality, of music, 775
 fluency heuristic, 339
 definition of, 339
 information processing influenced by, 339
 for moral judgment, 375
 fluid intelligence, 84
 fMRI. *See* functional magnetic resonance imaging
 fNIRS. *See* functional near infrared spectroscopy
 folk biology, concepts in, 198–199
 interdomain differences in, 199
 reasoning influenced by, 199
 foresight, thinking and, 1
 forgetting fixation theory, 459–461
 incubation effects in, 459
 reminiscence in, 459–461
 formal operations theory, 118
 formal thought disorder, 675. *See also*
 thought disorders
 activation spreading in, 681–682
 capacity allocation in, 683–684
- cognitive deficits and, integration of, 684
 communication deviance, 690–691
 context attention in, 682–683
 as etymologically precise labeling, 676
 excessive yielding to normal biases in, 683
 historical distinctions for, 675
 idea disassociation in, 675
 information processing deficits in, 681–684
 in new populations, 691
 normal speech production model and, 676–678
 ratings of, 675
 selective attention in, 682–683
 semantic memory retrieval deficits in, 682
 speech abnormalities as, 675
 word substitution in, 678
 forward-driven medical reasoning, 740
 Fosbury, Dick, 496
 4-term analogies, 236
 frame of mind, decision making and
 drive, 314–315
 emotion, 314–315
 identity and, 314
 metacognitive influences, 315
 priming and, 314
 framing effect, aging and, 655
 Frederick the Great, 494
 Freud, Sigmund, 3, 497
 unconscious thought for, 3
 frontal lobes. *See* prefrontal cortex, of brain
 frontotemporal lobar degeneration (FTLD), 72–73, 86
 symptoms of, 72–73
 Frost, Robert, 238
 FTLD. *See* frontotemporal lobar degeneration
 functional fixedness, 417
 functional magnetic resonance imaging (fMRI), 73, 78–80
 advantages of, 78
 brain networks under, 79–80
 for CH game model, 355–356
 gesture under, 637
 for insight activation, 476–478
 mental number line under, 591–592
 multiple memory systems under, 186
 numerical symbol development under, 595
 of PBC game, 355
 functional near infrared spectroscopy (fNIRS), 80–81
 for belief-bias effect studies, 81
 functional optical imaging, 80–81
 EROS, 81
 fNIRS, 80–81
 fundamental attribution error, 733
 future research applications
 for argumentation, 292–293
 for Bayesian inferences, 33

future research applications (*cont*)
for business, reasoning in, 769
for business schools, goals for, 769
for cognitive neurogenetics, 103, 104
for computational models, for concepts, 202
for connectionist neural-network models, 64
for decision making, 316
for dual-process theories, for deductive reasoning, 129–130
for game theory, 359–360
for gesture, in thought, 644
for heuristics and biases approach, to judgment, 341
for inductive inferences, 33
for learning from others, among children, 523–524
for mental model theory, for deductive reasoning, 151
for motivated thinking, 403–404
for neuroimaging, 85
for non-symbolic quantitative thinking, 602
for normative theories, of reasoning, 18–20
for problem solving, 428–429
for process models, of higher cognition, 64
for similarity assessment, 169–172
for symbolic-connectionist models, 64
for symbolic process models, 64
for symbolic quantitative thinking, 602
for visuospatial thinking, 624
fuzzy set theory, 182

G
Galton, Francis, 492–493. *See also* genius
game theory. *See also* Nash equilibrium
as abstract, 346
alternating-offer bargaining in, 353–355
asymmetric matching pennies game, 351
behavioral, 359
belief-dependent utility for, 360
CH model for, 347–348, 348–358
components in, 347
definition of, 347
dominance solvable games, 351
entry games in, 351–352
epistemic requirements, 347
equilibrium points in, 347, 360
function of, 346–347
future research applications for, 359–360
insula imaging and, 355
PBC games, 348–351
PG model for, 346, 348, 358–359
player knowledge as part of, 347
precuneus imaging and, 355
rational choice theory and, 18
strictly dominant strategy in, 349
weakly dominant strategy in, 349–350
Gandhi, Mahatma, 497
Gazzaniga, Michael, 68
gender
genius and, 505
wayfinding by, 617
general intelligence perspective, genius and, 496–499
additional intelligences in, 497
advantages of, 496–497
intelligent personality in, 497
limitations of, 497–498
General Problem Solver (GPS), 419
means-end analysis with, 419, 420
subgoals with, 419, 419–420
general slowing theory, 651
limited time in, 651
simultaneity in, 651
The General Theory of Employment, Interest and Money (Keynes), 348–349
generic memory, 201. *See also* semantic memory
generic noun phrases, 193–194
Genetic Studies of Genius (Terman), 493
genius
blind variation perspective for, 501–504
during childhood, 495
contemporary *vs.* posthumous recognition of, 494
within correlational psychology, 505
CQ test for, 495–496
definitions of, 492, 493–496, 496, 504
domain expertise perspective for, 498–499
etymology of, 493
exceptional leadership as, 494
future research applications on, 504–505
gender and, 505
general intelligence perspective for, 496–499
heroes as, 494
heuristics perspective for, 499–501
historic research on, 492, 492–493, 505
historiometric methods for, 493
for Kant, 494
lab experiments for, 493
magnitudes of, 494
nature *vs.* nurture debate for, 498–499
outstanding creativity as, 494
phenomenal achievement as, 493–494
prodigious performance and, 495
savants and, 495
within scientific psychology, 504–505
superlative intellect and, 495–496
theory of relativity as, 494
wisdom and, 665
geometric models, in similarity assessment, 155, 157–160

abduction reasoning and, 163, 164
augmentation of, 161
featural models and, 163–165
hierarchical cluster analysis in, 163–164
hierarchical representations in, 164
limitations of, 160, 161
propositions in, 164
unstructured representations in, 164
Gestalt psychology, 7
candle problem in, 417
functional fixedness in, 417
grouping principle, 426
in heuristics and biases approach to judgment, 327
insight in, 475, 478, 479
insight problems in, 423
for mathematical problem solving, 417–418
nine-dot problem in, 417
perceptual principle, 426
problem isomorphs in, 420–421
in problem solving, 3, 416–418
representation formation, for problem solving, 421
solution searching in, 420–423
two ropes problem in, 417
gesticulation, 632
gesture, in thought, 631–632
action roots for, 642–643
in adults, 633
affect displays, 632
autism and, diagnosis of, 635
in children, 633
cognitive development and, 633
communication through, 636–638
conceptualization demands for, 639
conceptual knowledge through, 633
cross-linguistic differences in, 641
deictic, 639
EEG imaging for, 637–638
emblems in, 632
evolutionary time span for, 641–642
under fMRI, 637
functions of, 632, 636–640
future research applications for, 644
ideation from, 640
illustrators, 632
intention of, 638
internal representations and, 639–640
knowledge activation through, 640
language development and, 633–634, 634–635
learning and, 634–636, 644
lexical access facilitation, 638–639
for listener, 636–638
mathematical equivalence through, 633, 636
motor imagery and, 642–643
as nonverbal behavior, 631
for persistent delay diagnosis, 635
potential function of, 631
production of, 640–643

- prosodic level of, 641
 regulators, 632
 self-adaptors, 632
 for speaker, 638–640
 speaker knowledge and, 631
 without speech, 637
 speech roots for, 640–642
 synergistic effects of, 644
 at syntactic level, 641
 task mastery and, in children, 635–636
 teaching applications for, 637
 Tower of Hanoi problem and, 632, 643
 translation into speech, 636–637
 as unspoken communication, 632–634
 visuospatial metaphors and, 642
 visuospatial thinking and, 639, 642, 644
 working memory and, 639
- Giere, R., 704
 Gigerenzer, G., 128
 Gladwell, Malcolm, 128
 global perspectives, on decision making, 311–315
 for decision framing, 311
 for decisions regarding future, 312–313
 frame of mind and, 313–315
 for future events, 313
 for mental accounting, 312
 for repeated decisions, 311–312
 for temporal discounting, 312–313
 goal errors, 746–747
 heuristics as source of, 746–747
 goal paths, in inhibit deficit theory, 652
 goal planning, categorization for, 178
 Goldberg, E. (Goldman), 665
 Gooding, D., 704
 Goodman, Nelson, 171
 Gorman, M.E., 704
 Gould, Stephen Jay, 514
 GPS. *See* General Problem Solver
 Graham, Martha, 497
 grammar theory, 372–373
 graphical displays, 620–623
 categorization of, 620
 for cognitive enhancement, 620
 comprehension of, 621–623
 format choice for, 621
 iconic, 620
 individual knowledge of, 623
 information layout in, 622–623
 interactions with, 623
 production of, 620–621
 relational, 620
 graph matching, 53
 Great Rationality Debate, 433, 435–436
 individual differences in rationality, 437–438
 irrationality in, 434
 Meliorists in, 435–436, 449
- Panglossians in, 436, 447
 group cognition, 763
 grouping principle, 426
 for objects, in language, 551
 for substances, in language, 551
 group polarization, 574
 guided activation model, 679–681
 prefrontal cortex role in, 679–680
 GUIDON program, 738
 guilt
 as moral emotion, 369–370
 in PG model, 346, 358
Gut Feelings (Gigerenzer), 128
- H**
- Handbook of Consumer Psychology*, 756
 hard cases, legal reasoning for, 722–724
 decisions for, 724–725
 multiple and inconsistent rules in, 723–724
 selection effect in, 724–725
 unclear rules in, 722–723
 wrong answers from rules in, 724
 harm/care domain, in moral judgment, 365
 harmonicity, 783–784
 harmony, in music, 776
 melody and, 776
 Hartmann, Stephan, 292
 HBM. *See* Hierarchical Bayesian models
 Heinz dilemma, 366
Hereditary Genius (Galton), 492–493
 heroes, as geniuses, 494
 Hesse, Mary, 238
 heuristics
 attribution substitution in, 373
 Bayesian inferences and, 25, 26
 bias-variance tradeoff with, 27
 catalog of, 374
 classification model, in medical reasoning, 739
 for emotion, decision making and, 315
 goal errors from, 746–747
 intention errors from, 746–747
 for judgments, 322–323
 in moral judgment, 373–374
 in moral reasoning, 373
 for physicians, 373–374
 in problem solving, 419
 for sacred and protected values, 383
 heuristics, for genius assessment, 499–501
 advantages of, 500–501
 BACON program, 500–501
 individual difference variables, 501
 limitations of, 501
 research applications for, 500–501
 Simon on, 500
 heuristics and biases approach, to judgment, 322–323
 aspiration levels in, 325
 attribution substitution account for, 340
- bounded rationality in, 324, 327
 causal base rate issue in, 336
 cognitive illusion paradigm in, 327
 conjunction fallacy and, 337
 critiques of, 326
 development history for, 341
 under dual-process theory of reasoning, 328
 future applications for, 341
 Gestalt psychology as influence on, 327
 historical antecedents to, 324–326
 intuitive conceptions in, 325
 logic of, 327
 negative model for, 333–336
 perception principles in, 326
 positive model for, 328–333
 probability in, 326
 prospect theory and, 327
 psychological structure of, 328
 psychophysics principles in, 326
 satisficing in, 325
 signal detection model and, 325
 surface structure of, 328
 visual illusion paradigm in, 326–327
- Hierarchical Bayesian models (HBM), 29–30
 for abstract causal knowledge theories, 30, 30
 for empiricism, 30
 for language, 30
 for nativism, 30
 hierarchical cluster analysis, 163–164
 high-conflict dilemmas, 370
 hints
 implicit, 485
 for insight, 485
 hippocampal volume, schizophrenia and, 688–689
 histones, 94
 Hobbes, Thomas, 3
 Holmes, Oliver Wendell, 725
 horizontal precedent, in legal reasoning, 727
How to Solve It (Polya), 416
 humans
 prosocial tendencies among, 533, 534, 535
 rationality and evolution for, 436, 436–437, 437
 human mind
 brain size as influence on, 530–531
 cognitive ability of, 529
 cognitive uniqueness of, 530–531
 cultural influence on, 533–534
 future research applications for, 537–538
 historical research on, 529
 language development and, 531–532, 533
 Machiavellian intelligence of, 530

- human mind (*cont*)
- module hypothesis for, 535–536
 - nonhuman minds compared to, 529–530
 - prosocial tendencies as influence on, 533, 534, 535
 - relational reinterpretation hypothesis for, 536–537, 537
 - role-based relational reasoning and, 536–537, 537
 - in social brain hypothesis, 530
 - in theory of mind, 534–535
- Human Problem Solving* (Newell/Simon), 418, 756
- human society. *See* society, human
- Hume, David, 3
- causal learning for, 213
 - moral judgment for, 368–369
- Hutcheson, Joseph, 723
- hypermnesia, 459
- hypotheses testing, 703–704, 705–706
- confirmation bias in, 705
- hypothesis-driven medical reasoning, 740, 740
- hypothetico-deductive model, in medical reasoning, 737
- I**
- iconic displays, 620
- production of, 621
- iconicity, principle of, 136–137
- advantages of, 137
 - visual images and, 136
- idea dissociation, in formal thought disorder, 675, 675–676
- idea roadmaps theory, 467
- fixation in, 467
 - path-of-least-resistance theory and, 467
 - problem solving and, 467
- ideation, 465
- brainstorming and, 465
 - convergent thinking and, 465
 - divergent thinking and, 465
 - from gestures, 640
- identical knowledge components theory, 800
- identity, decision making and, 314
- tension from, 314
- ignorance, arguments from, 282, 283–284
- Bayes theorem for, 284, 285
 - inferences from nothing in, 286
 - pragmo-dialectical theories for, 283
 - silence in, 285
- IGT. *See* Iowa Gambling task
- ill-defined problems, 463
- illusory inferences, in deductive reasoning, 142–143, 143
- principle of truth and, 142, 142
- illustrators, 632
- imagery. *See* visual imagery
- imaging genetics. *See* cognitive neurogenetics
- imagining objects, in visuospatial thinking, 608–611
- attention allocation for, 611
 - characteristics of, 608–609
 - in complex spatial tasks, 611–613
 - interaction of hands in, 610–611
 - mental precepts and, 609
 - mental rotation strategies for, 610, 610–611
 - nature of representations in, 609
 - neural basis for, 609–610
 - neuroimaging for, 609
 - object interaction and, 610–611
 - tacit knowledge in, 609
 - transformation of, 610
 - visual imagery in, 608–612
- Imhotep, 492
- imitation, as learning tool, 522
- cultural transmission and, 572
 - emulation compared to, 522
 - exaggerated, 522
 - of intentional actions, 522
- impasses, in insight problems, 483
- implicit cognition, creativity and, 463–464
- below-threshold activation in, 463–464
 - intuition and, 463
- implicit hints, 485
- implicit knowledge, for spatial tasks, 612–613
- implicit learning, 803. *See also* non-language-mediated learning
- from categorization, 186
 - neuroimaging for, 186
- implicit memory, creativity and, 459, 463–464
- for music, 778
- inconsistency, in normative reasoning theories, 19
- paraconsistent logics, 19
- inconsistent equations, 425
- incremented difficulty approach, 664–665
- incubation, of insight problems, 483
- incubation effects, of memory, 459
- forgetting fixation theory and, 459
 - TOT resolution and, 461–462
- individual differences
- in CH model, 355–356
 - in connectionist neural-network models, with mapping, 58
 - in developmental thinking, in children, 517
 - in heuristics, for genius assessment, 501
 - in organizational behavior, biases in, 760
 - for spatial layout knowledge learning, 617
 - in visuospatial thinking, 619–620
- individual differences, in rationality, 437–438, 448–449
- for Panglossians, 438
 - for tripartite theory of rationality, 441–445
- individual knowledge, 570
- induction, 145–147
- aging and, 658–659
 - analogy and, 235
 - categorical, 707
 - conditionals in, 146, 146–147
 - definition of, 145
 - essentialism and, 196
 - explicit, 146
 - inferences in, 22
 - intuitive, 146
 - as knowledge-dependent, 146
 - model representations of, 147
 - modulation principle in, 146
 - of schemas, 251–252
 - in scientific thinking and reasoning, 707
 - semantic information in, 145–146
- inductive inferences, 2, 22. *See also*
- Bayesian inferences
 - abstract learnability questions, 29
 - applications for other disciplines, 31–33
 - with Bayesian methods, 22–23, 29
 - causal maps for, 28
 - cognitive development and, 29
 - constraints on, 29–30
 - in developmental psychology, 32–33
 - explanations and, 271
 - future applications for, 33
 - HBM for, 29–30
 - iterated learning and, 29
 - origins of, 28–30
 - prior distributions in, 29
 - symbolic representations and, 30
- inductive reasoning, in medical reasoning, 738
- infants
- arithmetic for, 592–593
 - auditory sequence discrimination by, 587
 - cross-modal mapping by, 588–589
 - event representation for, 518–519
 - magnitude processing by, 589–592, 601
 - mental number line origins in, 589
 - non-numerical cues for, 587–588
 - number representation by, 587–589, 588–589
 - numerical discrimination by, 586–587
 - numerical expectations of, 592–593
 - numerical quantities for, 587–588
 - object representation for, 518–519
 - quantitative thinking for, 586–593
- inferences. *See also* analogical inference; Bayesian inferences; causal learning
- abductive, 268
 - background knowledge and, in problem solving, 428
 - category learning and, 183, 186–187
 - causal, 215–218
 - from concepts, 458–459
 - conditional, 118, 122–123

- in deduction, 2
 deductive, explanations and, 271
 explanations and, 268–270
 illusory, 142–143
 inductive, 2, 22
 intentions as, 382
 invalid, in deductive reasoning, 141–142
 language and, 546
 in medical reasoning, 737
 in mental model theory, 139–140
 in principle of truth, 139–140
 in rational models, 324
 in reasoning, 2
 inference learning, 183, 186–187
 inferential concepts, 192–193
 compositionality in, 192
 comprehension and interpretation in, 193
 information cascades, 573–574
 information processing
 context attention in, 682–683
 in directional outcomes, 393–394
 in dual-process theories, for deductive reasoning, 115, 116
 fluency heuristic and, 339
 in formal thought disorder, 681–684
 in medical reasoning, 737
 in mental representations, 37
 in nondirectional outcomes, 396
 selective attention in, 682–683
 semantic memory retrieval deficits in, 682
 in strategy-motivated thinking, 402, 402–403
 with technology, 750
 information systems, 764
 in-group/loyalty domain, in moral judgment, 365
 inhibit deficit theory, 652
 attentional resources in, 652
 goal paths in, 652
 inhibitory cues, 216
 inquiry learning approach, in science education, 713–714
 insight. *See also* insight problems
 “Aha” reaction, 476
 analytic solving and, 475, 478
 brain regional activation for, 476
 conscious analysis of, 478, 478–479
 through counterfactual mindsets, 488
 CRA problems and, 476
 creativity and, 487–488
 current perspectives in, 486–487
 definition of, 475
 EEG imaging for, 476–478
 enhancement of, 488
 ERPs and, 480
 fixation and, 480
 under fMRI, 476–478
 future research applications for, 487–488
 in Gestalt psychology, 475, 478, 479
 hints for, 485
 historical research on, 476
 knowledge selection and, 479–481
 methodological developments for, 486, 486–487
 neurofacilitation of, 488
 neuroimaging studies of, 487
 NRT and, 480
 positive emotion and, 484
 in prefrontal cortex, activation of, 477
 after preparatory brain state, 485
 problem solving and, 462–463, 475–478
 productive thought in, 479
 in real-world innovations, 487
 at resting brain state, 485–486
 unconstrained hypothesis generation for, 478
 verbal overshadowing of, 478–479
 warmth ratings and, 464
 insight experiences, 462
 insight problems, 423, 423–424
 anagrams as, 482
 attentional resources for, 484
 classic, 486
 fixation in, 483–484
 in Gestalt psychology, 423
 impasses in, 483
 with implicit hints, 485
 incubation of, 483
 neural activity during, 482
 nine-dot problem as, 417, 479
 progress monitoring theory for, 481
 radiation problem, 479–480
 representational change theory and, 423, 481
 restructuring of, 481–483
 solving of, 462–463, 475–478
 training intervals for, 484–485
 working memory in, 483
 instructional strategies, for knowledge transfer, 796–797
 goals of, 796
 instrumental rationality, 435
 insula, in brain, 355
 intellect, genius and, 495–496
 IQ tests, 495–496
 measurement of, 495–496
 intelligence. *See also* general intelligence
 perspective, genius and; genius CHC theory of, 443
 cognitive bias and, 122
 crystallized, 84, 443
 early studies of, 3
 fluid, 84
 frontal lobes and, role in, 84–85
 genius and, 496–499
 IQ tests, 443–444
 Machiavellian intelligence, 530
 rationality and, 438
 in tripartite theory of rationality, 443, 443
 type 2 processes and, in dual-process theory, 122
 visuospatial thinking and, 606
 intelligence quotient (IQ) tests, 443–444
 false outcomes from, 497
 genius level on, 495–496
 reliability of, 495
 intelligent personality, 497
 intensional reasoning, 144
 intensive care unit, decision making in, 747
 intention, moral judgment and, 380–382
 DDE and, 380
 definition of, 380–381
 as inferences, 382
 moral luck and, 381
 outcome bias and, 381
 side-effect effect and, 381–382
 intentional actions, imitation of, 522
 intention errors, 746–747
 heuristics as source of, 746–747
 intermediate effect, in medical reasoning, 743–744
 task recall in, 744
 intermediate phenotype approach, in cognitive neurogenetics, 92
 intermediate phenotypes, in schizophrenia, 688
 internal representations, 424–425, 639–640
The International Handbook of Research on Conceptual Change, 710
 intervention, 213
 in Bayesian inferences, 226–227
 in causal inference, 216, 217–218, 226–227
 observation compared to, 217–218
 placebo effects and, 216
 intuition
 in creativity, 464
 implicit cognition and, 463
 warmth ratings and, 464
 intuitionist theories, for moral judgment, 367–370, 384
 social institutionist model, 367–368
 intuitive induction, 146
 intuitive judgment, 128, 128–129
 in heuristics and biases approach, 325
 support theory and, 338
 invalid inferences, in deductive reasoning, 141–142
 Iowa Gambling task (IGT), 654–655
 IQ tests. *See* intelligence quotient tests
 irrationality, 12
 cognitive science and, 434
 in Great Rationality Debate, 434
 mindware and, 446
 isolation costs, 574–575
 rapid population change and, 574–575
 iterated learning, 29
 iTunes, 576

J

James, William, 41, 169
Jansky, Karl, 498
JDM view. *See* judgment and decision making view
Jen wu chib (The Study of Human Abilities) (Liu Shao), 505
Johnson, Samuel, 496
Johnson-Laird, Philip, 3
JOLs. *See* judgments of learning
judges
 analogical reasoning by, 731
 authority of law and, 731–732, 731–732
 expertise of, 731–732
 fact finding by, 730–731
 precedent for, 727
 unclear rules for, 723
 on US Supreme Court, 733
judgments. *See also* heuristics and biases
 approach to judgment; moral judgment
 aging and, 653–656
 attribution substitution account in, 340, 340–341
 for behavioral finance, 765–767
 from causal learning, 230
 conjunction fallacy for, 337
 dual-process theories for, 339–341
 explanations and, 270
 fluency heuristic for, 339
 framing effect and, with aging, 655
 heuristics research for, 336–337
 IGT and, with aging, 654–655
 intuitive, 128, 128–129
 motivational factors for, with aging, 655
 planning fallacy for, 336
 properties of, 2
 psychology of, development history of, 4
 ratio bias in, 340
 rational model for, 323–324
 similarity assessment and, 170
 support theory and, 338–339
 thinking and, 1
judgment and decision making (JDM) view, 26, 27
 for Bayesian inferences, 25–26, 26, 27
 for moral judgment, 374
judgments of learning (JOLs), 662–663
juries
 burden of proof, 729
 decision making by, psychological model of, 730
 deliberation instructions for, 729
 explanatory coherence model for, 730
 fact finding by, 729–730
 facts for, 721–722
 story model for, 730
 trial instructions for, 729

K

Kahneman, Daniel, 4, 326, 339–340. *See also* heuristics and biases approach, to judgment
 on conjunction fallacy, 337
 entry games for, 351
 judgment heuristics for, 322–323
 visual attention for, 339–340
Kant, Immanuel, 3
 genius for, 494
 moral judgment for, 368
Keller, Helen, 543
Kepler's Third Law of Planetary Motion, 500
Keynes, John Maynard, 348–349
Kincannon, A., 704
Kinds of Minds (Dennett), 441
Klahr, D., 704
KLI framework. *See* knowledge-learning-instruction framework
Knobe effect, 381–382
knowledge
 of basic science, in medical reasoning, 742, 745
 in behavioral accounting, transfer prediction for, 768
 clinical, in medical reasoning, 742, 745
 compiled causal, in medical reasoning, 746
 in creativity, cognition and, 457, 469–470
 individual, 570
 insight and, 479–481
 medical, 738, 742, 742, 743–744
 organizational structure of, 458
 in representation of culture, mediation of, 572
 tacit, 609
 in transmission of culture, mediation of, 572
knowledge activation
 in directional outcomes, 394–395
 through gesture, 640
 in nondirectional outcomes, 396–397
knowledge effects, in problem solving, 424–428
 background, 427–428
 inferences in, 428
 organization of, in medical reasoning, 738, 742, 743–744
 semantic alignment and, 428
knowledge encapsulation, in medical reasoning, 745
knowledge-learning-instruction (KLI) framework, 791
knowledge representation
 anti-representationalism in, 48–49
 in cognitive psychology, 47
 content in, 47
 in developmental psychology, 47
 embodied cognition in, 47–48
 featured representations, 42–43
 knowledge transfer and, 802–803
mental representations, 36–39
situated cognition in, 48
spatial representations, 39–42
structured representations, 43–44
structure in, 47
knowledge transfer
 with ACT-R, 793–794
 of algebra, as symbolic language, 798, 803, 803
big ideas as part of, 791–793, 798
in cognitive load theory, 797
components of, 791–793
component testing of, 793–794
computational models for, 793–794
difficulty factors assessment for, 799 with domain-general approaches, 795–801
with domain-specific approaches, 795–801
educational data mining for, 799–800
element discovery for, 798–801
empirical analysis for, 798–799
identical knowledge components
 theory of, 800
instructional strategies for, 796–797
with KLI framework, 791
knowledge-based dependencies in, 797–798
knowledge representation and, 802–803
learning to think as, 789
methods for, 798–801
modularity of, 793–794
in pieces, 789
scope of, 789
small ideas in, 791–793, 798
for symbolic languages, 789
vertical transfer in, 793
Köhler, Wolfgang, 3
Krumhansl, Carol, 161

L

labeling, in language in, 562
 for color, 549–550, 563
labor. *See* cognitive division of labor
landmarks, wayfinding with, 617
language
 in alignment-based models, structural accounts with, 167
 arbitrariness of meaning in, 547–548
 Broca's area and, 68
 categorization and, 189, 545
 child development of, gesture and, 633–634, 634–635
 in children's developmental thinking, 513–514
 code-switching in, 561
 cognitive neurogenetics for, 102–103
 color perception in, 549–550
 concepts and, 189
 core knowledge of, 545
 cultural definitions with, 543
 dedicated neural systems for, 79
 development of society through, 577

- direct effects of, 548
 dishabituation paradigm for, 548
 empiricist perspective on, 544
 for evidentiality, 555–556
 explanations and, 272
 generic noun phrases, 193–194
 genetic polymorphisms and, 102
 genotypic mutations and, 102
 gesture and, 633–634, 634–635
 HBMs for, 30
 human cognition and, 531–532, 543
 human mind development and, 531–532, 533
 inconsistent encoding in, 547–548
 indirect effects of, 548
 inference and, 546
 as innate, in humans, 532
 interpretive flexibility of, 548
 iterated learning for, 29
 labeling in, 562
 language-on-language effects and, 562, 563
 learning to think with, 794–795, 801
 linguistic intrusion and, 562
 linguistic-relativistic view of, 544–545
 memory and, for word meaning, 181
 in module hypothesis, 536
 for motion, 553–554
 multiple interpretations within, 546–547
 for number, 557–558
 for objects, 550–552
 permanent effects of, 548
 phonetic restructuring of, 548–549
 polysemy, 194–195
 priming, with memory, 181
 relational, in mapping, 247
 relativity of, 548, 549
 reorganization of thought and, 548–561
 richness of, compared to thoughts, 547
 similarity tests for, 563
 for spatial frames of reference, 554–555
 for spatial orientation, 558–561, 563
 for spatial relationships, 552–553
 for substances, 550–552
 symbolic, 789
 thinking and, 1, 801–802
 as thought, 544–546, 546–547
 for time, 556–557
 transient effects of, 548
 universalist perspective on, 544
 varieties of, 532–533
 Whorf-Sapir hypothesis for, 544–545, 561, 562
 language-on-language effects, 562, 563
 language-on-language interpretation
 for objects and substances, 552
 for spatial relationships, 553, 552–553
 law, argumentation in, 279, 291–292, 293
 burden of proof in, 279
 dialectical model of, 279
 facts and, distinction between, 721–722
 pragma-dialectical theories of, 279
 testimony and, 292
 leadership, as genius, 494
 learning. *See also* causal learning; feature learning; implicit learning; inference learning; iterated learning; learning from others, among children
 in behavioral finance, 767
 from categorization, 178, 179
 in consumer behavior, 757
 domain specificity in, 198
 explanation in, 265–268
 gesture and, 634–636, 644
 intervention in, 213
 learning to think compared to, 789
 in marketing, 757
 in organizational behavior, 760–761
 PBL, 748–749
 Learning and Inference by Schemas and Analogies (LISA), 61–62, 82–84
 analogical reasoning with, 241, 241
 analog retrieval by, 241
 behavioral data simulations by, 241
 mapping connections in, 241
 learning by doing principle, 797
 learning from others, among children, 521–524
 benefits of, 521–522
 flexibility of, 523
 future research on, 523–524
 through imitation, 522
 pedagogical goals in, 522
 through social interaction, 524
 through testimony evaluation, 523–524
 learning from worked-out examples principle, 797
 learning systems, in management science, 764–765
 delayed feedback in, 764–765
 learning to think, as knowledge transfer, 789
 components of, 790–791
 language development and, 794–795, 801
 with NLM learning, 801
 with symbolic languages, 790
 legal categories, 725–726
 legal decisions, 733
 for hard cases, 724–725
 legal procedures, 732–733
 structural differences in, 732
 Legal Realists, 720–721
 unclear rules for, 723
 legal reasoning. *See also* judges; juries
 analogical reasoning in, 727–728
 as artificial reason, 719
 authority of law in, 731–732, 731–732
 categorization in, 725–726
 in civil law nations, 733
 in common law nations, 733
 confirmation bias and, 720–721
 decisions under, 733
 definition of, 720
 distinction within legal system, 720
 under English common law, 723–724
 explanatory coherence model in, 730
 fact finding and, 728–731
 fundamental attribution error in, 733
 for hard cases, 722–724
 law and fact in, distinctions between, 721–722
 legal categories in, 725–726
 legal procedures and, 732–733
 Legal Realists and, 720–721
 moral intuitionist model for, 733
 parameters of, 720
 precedent in, 726–727
 relational categories for, 726
 rules for, 722–725
 selection effect in, 724–725
 in stare decisis, 727
 story model in, 730
 legitimacy effects, on directional outcomes, 393
 lemma retrieval, 677
 speech disorders and, 678
Leviathan (Hobbes), 3
 Lewis, C.I., 211
 linear presentations, in symbolic arithmetic, 599
 linear reasoning, in visuospatial thinking, 619
 linguistic intrusion, 562
 linguistic-relativistic view, of language, 544–545
 cues in, 545
 LISA. *See* Learning and Inference by Schemas and Analogies
 Liu Shao, 505
 localist representations, in connectionist neural-network models, 56
 localizationism theory, 68
 local perspectives, on decision making, 311–315
 for decision framing, 311
 for decisions regarding future, 312–313
 frame of mind and, 313–315
 for future events, 313
 for mental accounting, 312
 for repeated decisions, 311–312
 for temporal discounting, 312–313
 logic, 12–16. *See also* mental logics
 account of consequence in, 14
 in argumentation, 278
 belief and, 12, 18
 completeness of, 14
 contradictions in, 13, 13
 deontic, 15
 in epistemic argumentation, 282
 Fermat's Last Theorem and, 13

- logic (*cont.*)
- of heuristics and biases approach to judgment, 327
 - languages, 13
 - local consistency in, 12, 14–15, 18
 - modal, 15
 - in normal speech production model, 677
 - paraconsistent, 19
 - precision of intuitions in, 13
 - psychological role of, 15
 - as semantically valid, 14
 - soundness of, 14
 - as syntactically valid, 14
 - temporal, 15
 - thinking and, 2
 - truth functional connectives in, 147
- logic problems, 3
- long-term memory
- analogical reasoning in, 244–245
 - BDNF gene in, 100–101
 - cognitive neurogenetics for, 100–102
 - concepts in, storage of, 201
 - dopaminergic genes in, 101
 - genetic polymorphisms in, 101–102
 - genotype scores in, 101
 - neurotransmitter-related polymorphisms in, 101
 - relational structure in, 244–245
 - retrieval gaps in, 244
 - schizophrenia and, 686, 689
- loss aversion, 305–306
- endowment effects and, 305, 306
 - research on, 306
 - status quo and, 305–306
 - trade reluctance in, 305
- luck. *See* moral luck
- Luther, Martin, 494
- M**
- Machiavellian intelligence, 530
- macropsychology
- culture and, 579–580
 - definition of, 569–570
- magical realism, 2
- magnetic resonance imaging (MRI), 74
- magnetoencephalography (MEG), 80
- under Faraday's Law, 80
- magnitude processing, 589–592, 601
- math proficiency and, 600
 - mental number line in, 589–590
- management, in business. *See*
- organizational behavior
- management science, 763–765
- groups under, 765
 - information systems in, 764
 - learning systems in, 764–765
 - research for, 765
 - support system design, 763–764
- mapping, in analogical reasoning, 46, 235, 246
- alignability in, 246–247
 - analogy goals as influence on, 247
- attentional focus in, 246–247
- bistable, 247
- coherence in, 247
- complexity in, variability of, 249
- computational level analysis of, 253
- cross-mapping, 247
- cross-modal, by infants, 588–589
- developmental changes in, 248–249
- in LISA, 241
- relational language in, 247
- relational shift and, 248
- working memory and, 248
- marker passing, 44
- marketing, in business, 756–759
- analogy research for, 759
 - brand extension in, 758
 - consumer behavior, 756–759
 - context for, 757
 - decision making in, 756–757
 - expertise in, 757
 - learning in, 757
 - preferences and, 758–759
 - research in, 758
 - similarity research for, 759
- Marr, David, 4
- analysis for, 4
 - computation for, 4
 - representation and algorithm for, 4
- matching pennies game. *See* asymmetric matching pennies game
- mathematical equivalence
- for children, 636
 - gesture and, 633
- mathematical problem solving, 417–418
- math proficiency, magnitude processing and, 600
- MDS. *See* multidimensional scaling
- meaning, in mental representations, 38
- grounding and, 38
 - from semantic networks, 38
 - sources of, 38
- means-end analysis, in problem solving, 419, 420
- measurements, numerical symbols for, 598
- mechanistic explanations, 264
- medical outcomes, status quo and, 306
- medical reasoning
- abductive models of, 739
 - abstraction in, 739
 - with AI models, 737–738
 - Bayes Theorem for, 737
 - causal reasoning as part of, 744–746
 - with clinical knowledge, 742, 745
 - clinical reasoning in, 740
 - compiled causal knowledge in, 746
 - cover and differentiate model in, 739
 - data-driven, 740–741
 - decision analysis approach in, 737
 - decision making in, 746–747, 747–748
 - deductive models of, 738
 - definition of, 736–737
 - early research on, 738, 737–738, 751
- in educational settings, 748–749
- errors in, 746–747
- forward-driven, 740
- future research applications for, 750–751
- goal errors in, 746–747
- GUIDON program for, 738
- heuristics classification model in, 739
- hypothesis-driven, 740, 740
- hypothesis testing phase in, 739, 739–740
- hypothetico-deductive model, 737
- inductive models of, 738
- inferential reasoning in, 737
- information processing approach in, 737
- in intensive care units, 747
- intention errors in, 746–747
- intermediate effect in, 743–744, 744
- knowledge encapsulation as part of, 745
- knowledge organization in, 738, 742, 743–744
- medical knowledge and, nature of, 742
- mental models in, 745–746
- models of, 738–739
- MYCIN program for, 738
- NDM as part of, 747
- NEOMYCIN program for, 738
- PBL in, 748–749
- PIP for, 737–738
- problem solving in, 737, 740–741
- with science knowledge, 742, 745
- select and test model in, 739
- similarity of cases and, 741–742, 741–742
- during task transitions, 747–748
- technology-mediated, 749–750
- medication management, with technology, 750
- Meehl, Paul, 326
- Meerkat Manor*, 533
- MEG. *See* magnetoencephalography
- meliorism, 435–436
- Meliorists, 435–436, 449
- melody, 775
- harmony and, 776
 - as psychological construct, 775–776
- memory. *See also* episodic memory;
- implicit memory, creativity and;
 - long-term memory; procedural memory, for music performance;
 - semantic memory; working memory
- aging and, 660–662
- analogical reasoning and, 245–246
- capacity of, theories for, 3
- collaborative, 571–572
- concepts and, 178–179
- creativity and, 459–462
- directional outcomes, 395
- domain specificity for, 200–201
- forgetting fixation theory and, 459–461

- fragmentation of, within semantic memory, 181–183
- hypernesia, 459
- incubation effects of, 459
- in lexical decision tasks, 181
- as meaning-based, in connectionist models, 201–202
- metamemory, 662–663
- multiple systems for, 185
- organization of, 181–182
- personal relevance and, by age, 661–662
- priming in, with words, 181
- reminiscence and, 459
- TOT resolution and, 461–462
- transactive, 571
- word meaning and, 181
- Mendel, Gregor, 494
- mental accounting
- in decision making, 312
 - preferences in, 312
- mental line number, EEG imaging for, 591–592
- mental logics, 12
- cognition as proof theory and, 15
 - deductive reasoning and, 118
- mental model theory, for deductive reasoning, 118–119, 134
- analogy in, 136
- artifacts in, 136
- in cognitive science research, 136
- conclusions in, guidelines for, 135
- conditional inference in, 119
- conjunction procedure in, 140
- deduction in, 136–140
- diagrams in, 135–136
- discourse in, 136
- foundations for, 134–135, 151
- future applications for, 151
- history of, 135–136
- iconicity in, 136–137
- inferences in, 139–140
- perceptions in, 134
- possibilities in, 137–138
- premise guidelines for, 135
- as proposition-based, 135
- truth in, 138–140
- mental number line
- arithmetic and, 592–593
 - in Dehaene–Changeux model, 590
 - EEG imaging for, 591–592
 - encoding of, 591
 - fMRI imaging for, 591–592
 - in magnitude processing, 589–590
 - neural basis for, 590–592
 - origins of, in infants, 589
 - in prefrontal cortex, 591
 - under Weber–Fechner law, 589
- mental precepts, visual imagery and, 609
- mental representations, 36–39
- academic developmental history for, 3
 - algorithmic level of, 37
 - analog, 39
- behavioral approach to, 36
- computational level of, 37
- definition of, 37–38
- elements of, 39
- implementational level of, 37
- information processing in, 37
- meaning in, 38
- of motion, language for, 554
- represented world in, 37
- representing world in, 37
- of spatial orientation, language for, 559
- states of endurance for, 39
- symbols in, 39, 39
- theoretical approach to, 36
- in thinking, 1–2
- types of, 39
- mental road maps, 781
- mental rotation strategies, 610, 610–611
- mere exposure effect, 778
- metacognitive influences, on decision making, 315
- aging and, 662–663
 - metamemory in, 662–663
- metamemory, 662–663
- JOLs and, 662–663
- metaphor
- analogy and, 236–238
 - psychology research on, 238
 - for time, in language, 556, 556
 - visuospatial, 618
- meter, in music, 776
- metrics, of creativity, 456, 467–468
- abstraction, 468
 - conformity in, 468
 - emergence, 468
 - outcome, 468
 - process, 468
- microeconomics, rational choice theory and, 18
- microgenetic studies, for number-line estimation, 596–597
- Miller, Earl, 679
- Miller, George, 68
- mind-reading, 576
- mindware, 445–446
- crystallized intelligence and, 446
 - irrationality and, 446
- minimality assumption, 160
- mini-societies, 572–573
- Miranda v. Arizona*, 727
- modal logics, 15
- model-theoretic consistency, 14
- modularity
- in developmental thinking, 516
 - of knowledge transfer, 793–794
 - in spatial relationships, 559
- modulation principle, in induction, 146
- module hypothesis, for human mind, 535–536
- for language, 536
- Mohammed (Prophet), 494
- monetary outcomes
- decision making and, 303
- risky decisions for, framing effects of, 303–304
- status quo and, 305–306
- monkeys, brain size for, 530
- Monte Carlo methods, for Bayesian references, 33
- moral cleansing behavior, 385
- moral/conventional distinction, in moral judgment, 375–377
- development of, 375
- empirical issues with, 376
- moral rules in, 375, 377
- norms in, acquisition of, 377
- signature conventional pattern in, 376
- signature moral pattern in, 375
- VIM in, 376
- moral dilemmas, 377–380
- emotion in, 380
 - preferred ethical position and, 380
 - test question for, 380
 - thought styles and, 380
 - trolley dilemma, 377–379
 - victim type and, 380
 - vividness of death and, 380
 - working memory and, 380
- moral emotions
- anger, 369
 - guilt, 369–370
- moral grammar theory, 371–373
- computational steps in, 372, 372
 - cultural bias in, 372
 - doctrine of double effect in, 372
 - foundations of, 371–372
 - grammar theory and, 372–373
 - as innate, 372
 - permissibility in, 371–372
 - trolley dilemma and, 377–378
- moral heuristics, 373–374
- catalog of, 374
 - critiques of, 374
 - “do no harm,” 373–374
- moral intuitionist model, for legal reasoning, 733
- morality. *See also* moral judgment
- conventional, 366
 - definition of, 376
 - postconventional, 366
 - preconventional, 366
- moral judgment
- attentional processes in, 375
 - authority/respect domain in, 365
 - causal model theory for, 374–375
 - cleansing behaviors and, 385
 - cognitive-affective theory and, 368
 - contemporary research on, 385
 - cultural influences on, 365,
 - 385–386
 - definition of, 364–365
 - dilemmas in, 377–380
 - domain-general cognitive theories, 374–375, 384
 - dual-process theory in, 370–371

- moral judgment (*cont*)
- emotion as influence on, 368–370, 384
 - fairness/reciprocity domain in, 365
 - fluency processing in, 375
 - harm/care domain in, 365
 - heuristics in, 373–374
 - Humean approach to, 368–369
 - in-group/loyalty domain in, 365
 - as innate, 365
 - intention and, 380–382
 - intuitionist theories for, 367–370, 384
 - Kantian approach to, 368
 - moral/conventional distinction in, 375–377
 - moral grammar theory in, 371–373
 - moral psychology and, 364
 - norms in, acquisition of, 365
 - purity/sanctity domain in, 365
 - rationalist theory of, 366–367
 - reasoning in, 365
 - rules acquisition in, 365
 - sacred and protected values and, 382–384
 - sentimental rules theory in, 369
 - social institutionist model for, 367–368
 - as social relations regulatory, 366
 - in Western paradigms, 385–386
 - moral luck, 381
 - moral psychology, 364
 - moral reasoning, 365. *See also* moral judgment
 - heuristics in, 373 - moral rules, 375, 377
 - conventional domains and, 375 - morpho-phonological encoding, 678
 - motion, language for, 553–554
 - attention allocation for, 554
 - cross-linguistic differences in, 553–554, 554, 563
 - mental representations of, 554 - motivated thinking
 - academic foundations of, 390–391
 - culture and, 577
 - directional outcomes in, 391, 391–395
 - emotion and, 404–405
 - future research applications for, 403–404
 - interactions among types of, 404
 - nondirectional outcomes in, 391, 395–397
 - outcome-based, 391–399
 - in prevention-focused individuals, 405
 - in promotion-focused individuals, 405
 - reasoning and, 404
 - strategy-based, 399–403 - motivational systems, for moral judgment, 375
 - motor imagery, gesture and, 642–643
 - Mouselab, 353, 360
- Mozart, Wolfgang Amadeus, 495
- MRI. *See* magnetic resonance imaging
- MSIT. *See* multi-source interference task
- multiconstraint theory, 239
- for coherence, in analogical mapping, 247
 - relational reasoning and, 239–240
- multidimensional scaling (MDS), 40–41
- bias in, 161
 - for compressed representations, 158–159
 - for object dimensions, 158
 - postulated representations from, 163–164
 - for quantitative representations, 159–160
 - in similarity assessment, 157–160
- multiple memory systems, 185
- under fMRI, 186
 - prototype extraction tasks for, 185–186
- multi-source interference task (MSIT), 105
- Murphy, Gregory, 177
- music
- attentional focus for, 784
 - auditory stream segregation for, 776
 - chords, 775
 - cognitive function influenced by, 779
 - composition, 782–784
 - definition of, 774–775
 - as discretionary, 774
 - dynamic attending in, 776
 - emotional response to, 778–779, 786
 - evolutionary perspective for, 785–786
 - expectancies in, 778–779
 - expertise in, 775
 - floating intentionality of, 775
 - function of, 774
 - future research goals for, 785–786
 - harmony in, 776
 - in implicit memory, 778
 - listening to, 775–776
 - melodies, 775
 - mere exposure effect for, 778
 - meter in, 776
 - neuroscience of, 786
 - passive exposure to, 785
 - performance of, 779–782
 - preferences for, 778
 - rhythms of, 775
 - sound attributes for, 776
 - structures in, 775, 776–778
 - tone in, 777
 - universality of, 775
- musical structure, 775
- through auditory scene analysis, 776
 - perception of, 776–778
 - through pitch perception, 776
 - sensitivity to, 777–778
 - tone in, 782, 784
- music composition, 782–784
- accessibility in, 782
- auditory perception principles for, 783
- cognitive constraints on, 782–784
- harmonicity in, 783–784
- psychophysical limits on, 783
- rules of voice-leading for, 783
- stylistic changes in, 782
- tonal structure in, 782, 784
- Mussweiler, Thomas, 333
- MYCIN program, 738
- Mynatt, C.R., 704
- ## N
- Nash, John, 347. *See also* game theory
- Nash equilibrium
- in asymmetric matching pennies game, 351–349
 - in entry games, 352
 - payoff predictions with, 350–351
- nativism, 30
- naturalistic decision making (NDM), 747
- natural selection, rationality and, 436
- The Nature of Explanation* (Craik), 135
- nature *vs.* nurture, for genius, 498–499
- navigation, through environmental space, 613
- allocentric representations in, 614
 - cognitive processes for, 613
 - egocentric representations in, 614
 - by gender, 617
 - processes in, 615–618
 - spatial representations in, 613–615
 - vision and, 613
 - wayfinding, 617–618
- NDM. *See* naturalistic decision making
- negative model, for heuristics and biases
- approach to judgment, 333–336
 - aggregate frequency in, 336
 - conservational perspective for, 335
 - critiques of, 335–336
 - evidential neglect from, 334, 334
 - frequency format for, 336
 - irrational judgments from, 334
 - probability judgments in, 335
 - strength-weight theory and, 334
 - validity issues from, 334–335
- neglect, in negative model for heuristics
- and biases approach to judgment, 334, 334
- NEOMYCIN program, 738
- neural systems. *See also* prefrontal cortex, of brain
- computational models of, 82–84, 85
 - for cultural cognition, 576–577
 - dedicated, 79
 - dedicated, for reasoning and language processing, 79
 - LISA, 82–84
 - for mental number line under, 590–592
 - for normal speech production model, 677–678
 - with schizophrenia, 685–689
 - for visual imagery, 609–610
- neurogenetics. *See* cognitive neurogenetics

- neuroimaging
 for arithmetic, with infants, 592
 cognitive neurogenetics and, 104
 in cognitive neuroscience, 73–81
 for concepts, in studies, 188–189
 cortical connectivity under, 85
 electrophysiological, 74–77
 for emotions, in moral judgment, 369
 functional optical imaging, 80–81
 future applications for, 85
 for gesture, 637–638, 637
 for implicit learning, 186
 integration of methods, 85
 mental number line under, 591–592
 numerical symbols under, 595
 for schizophrenia, 686
 spatial functional, 77–80
 structural, 73–74
 temporal functional, 85
 with TMS, 81–82
 for visual imagery, 609
- neuropsychology. *See also* cognitive neuropsychology
 emotion and, in moral judgment, 369
- neuroscience. *See also* cognitive neuroscience
 of music, 786
 social cognitive, 690
- New Directions in Scientific and Technical Thinking* (Gorman/Kincannon/Gooding/Tweneay), 704
- Newell, Allen, 3, 418, 418, 756
 problem solving for, 418–420, 704
 problem spaces for, 418–419
- Newton, Isaac, 492
- nine-dot problem, 417, 479
 progress monitoring theory for, 424
- NLM learning. *See* non-language-mediated learning
 nodes
 in spreading activation, 44–45
 in symbolic process models, 53
- nonconfounding conditions, in causal power theory, 220
- nondirectional outcomes, in motivated thinking, 391, 395–397
 accuracy for, 395, 405
 attribution effects on, 395–396
 closure motivations in, 397
 cognitive needs for, 405
 cognitive processes influenced by, 397
 evaluation complexity effect on, 396
 information search effects on, 396
 knowledge activation in, 396–397
 recall effects on, 396–397
- nonhuman minds, human compared to, 529–530
- non-language-mediated (NLM) learning, 801, 803
 educational design for, 802
- nonmonetary outcomes, decision making and, 303
- non-numerical cues, for infants, 587–588
- non-symbolic numerical processing, 593–594
 in arithmetic, 594
 internal Weber fraction in, 593
 numerical discrimination in, 593–594
 under Weber-Fechner law, 594
- non-symbolic quantitative thinking, 585, 586–594
 arithmetic, for infants, 592–593
 future research applications for, 602
 limitations of, 601
 magnitude processing, by infants, 589–592, 601
 mental number line in, 589
 numerical discrimination, by infants, 586–587
 processing in, 586
- nonverbal behavior, 631
- normal speech production model, 676–678
 context attention in, 682–683
 feedback loop in, 677
 lemma retrieval in, 677
 logic consistency in, 677
 at neural level, 677–678
 selective attention in, 682–683
- normative theories, of reasoning, 11, 11–12
 artificial intelligence and, 19
 Bayesian inferences in, 12
 complex calculations and, 19–20
 consistency conditions in, 12
 decision making and, 11–12
 descriptive theories and, 12
 future applications of, 18–20
 inconsistency and, 19
 irrationality and, 12
 in mental logics, 12
 rationality and, 12
 standards within, 11
- Nosofsky, Robert, 161
- notations, in symbolic process models, 52–53
- Not by Genes Alone* (Richerson/Boyd), 533
- noticing, in creativity, 469
- Novum Organum*, 702
- NRT. *See* number reduction task
- nucleotides, 91–92
 SNPs in, 92
- number, language for, 557–558
 categorization of, 598–599
 in children, 557
 counting systems and, 557–558, 563
 cross-linguistic differences in, 557–558, 563
- for large-number calculations, 557
- number words, 558
 quantifiers in, 558
 semantic properties for, 558
 symbolic quantitative thinking and, 585
- number board games, 600–601
- number-line estimation, 595–598
 analogies for, 598
 categorization in, 598–599
 developmental stages for, 596
 microgenetic studies for, 596–597
- number reduction task (NRT), 480
- number words, 558
- numerical cognition, for behavioral accounting, 768
- numerical discrimination, 586–587
 by infants, 586–587
 in non-symbolic numerical processing, 593–594
 with objects, 587
- numerical symbols, comprehension of, 594–601, 601–602
 categorization of, 598–599
 comparisons of, 595
 development patterns for, 595
 magnitude processing and, 600
 for measurements, 598
 under neuroimaging, 595
 number board games for, 600–601
 of number-line estimation, 595–598
 for numerosity estimation, 598
 symbolic arithmetic, 599–600
 theory-based interventions for, 600–601
 under Weber-Fechner law, 595
- numeric priming, 333
- numerosity estimation, numerical symbols for, 598

O

- object concepts, 188
 for infants, 518–519
 numerical discrimination for, 587
- object identity, in sорталism, 198
- objects, language for, 550–552
 categorization of, 551
 classificatory tasks for, 550–551
 grammatical determinants for, 550
 grouping of, 551
 language-on-language interpretation for, 552
 linguistic stimulus of, 552
 ontology for, 550
- objects, visuospatial thinking for, 607–613
 categorization of, 608
 imagining of, 608–611
 recognition of, 608
- observation, intervention compared to, 217–218
- “On Scientific Thinking” (Tweneay/Doherty/Mynatt), 704
- ontologies, in developmental thinking, 520–521
 for objects and substances, 550
- operational momentum, 593
- operations, for business. *See* management science
- operators, in problem solving, 414
- opportunistic assimilation theory, 466

- opportunistic assimilation theory (*cont*)
 prepared mind theory as part of, 466
 problem solving and, 466
- organizational behavior, 759–763
 categories in, 761–762
 cooperation in, 762
 creativity in, 762–763
 cultural influence on, 762, 762
 decision making for, 760
 ethics in, 760
 group cognition and, 763
 individual level biases in, 760
 learning in, 760–761
 research on, 761–762
- organization effects, in directional outcomes, 395
- orientation, in environmental space, 613
 allocentric representations in, 614
 after disorientation, 616
 egocentric representations in, 614, 615
 internal cues for, 615
 perspective taking for, 615
 processes in, 615–618
 spatial layout knowledge and, 616–617
 spatial representations in, 614–615
- outcome metrics, for creativity, 468
- outcome-motivated thinking, 391–399
 accuracy in, 398–399
 closure motivation in, 399
 directional, 391, 391–395
 limits to, 397–399
 nondirectional, 391, 395
 reality constraints to, 397–398
- own-age bias, 657
- P**
- Panglossians, 436, 447
 economic discipline for, 436
 extreme, 447
 individual differences in rationality for, 438
The Paper Chase, 719
- paraconsistent logics, 19
- parallel connectivity, in analogy, 46
- parallel distributed processing (PDP) models, 55. *See also* connectionist neural-network models
- parsimony
 in Bayesian inferences, 224
 in causal learning, 230
- path-of-least-resistance theory, 459
 idea roadmaps theory and, 467
- pattern recognition, 422
- PBC game. *See* p-beauty contest game
- p-beauty contest (PBC) game, 348–351
 as dominance solvable game, 351
 fMRI imaging of, 355
- PBL. *See* problem-based learning
- PDP models. *See* parallel distributed processing models
- Pearson, Karl, 213
- perception effects, in problem solving, 424–428
 with diagrams, 425, 426–427
 with equations, 425–426
 external representations in, 425
 internal representations in, 424–425
 visual, 424–427
- perception principles, 326
- perceptions, in mental models, 134
- perceptual fluency, 186
- perceptual-functional hypothesis, for category deficits, 187–188, 188
- perceptual learning modules, 425–426
- perceptual narrowing, 517–518
- perceptual principle, 426
- performance, genius and, 495
 domain expertise perspective and, 498
- performance, of music, 779–782
 expression in, 780, 780–781
 pitch in, 781–782
 procedural memory for, 780
 timing in, 781–782
 vocal control in, 779
- performance expression, 780, 780–781
 mental road maps for, 781
 procedural memory for, 780–781
- persistent delay, diagnosis of, 635
- PET. *See* positron emission tomography
- PG model. *See* psychological game framework
- phenotypes
 in cognitive neurogenetics, 92
 SNPs in, 92
- philosophy, argumentation in, 278
- phonetic encoding, 678
- physical symbol system hypothesis, 803
- physics, for children, 711
- Piaget, Jean, 3, 514. *See also*
 developmental psychology;
 developmental thinking, in children
 developmental stages for, 515
 formal operations theory, 118
- Picasso, Pablo, 497
- Pierce, Charles Sanders, 135
 abduction for, 268
- PIP. *See* Present Illness Program
- pitch, in music performance, 781–782
- pitch perception, 776
- placebo effects, 216
- Planck, Max, 501
- planning fallacy, 336
- pluralistic ignorance, 574
- Polya, George, 416, 791
- polygenicity, 92
- polysemy, 194–195
- positive feedback, for problem solving, 657
- positive model, for heuristics and biases
 approach to judgment, 328–333
 anchoring in, 328, 332–333
 availability heuristic in, 330–331
 cognitive illusion paradigm in, 329
 probability assessment in, 329
 regression artifacts in, 329–330
- representativeness in, 331–332
- positron emission tomography (PET), 77–78
 SPECT, 77
- possibilities, principle of, 137–138
 fully explicit models for, 137–138, 138–140
 sentential reasoning in, 137
- postconventional morality, 366
- posthumous recognition, 494
- Postman, Neil, 435
- power PC theory. *See* causal power theory
- PPA. *See* primary progressive aphasia
- practical problems, candle problem, 417
- pragma-dialectical theories, of
 argumentation, 279
 ignorance arguments, 283
- precedent, in legal reasoning, 726–727
 horizontal, 727
 for judges, 727
 vertical, 726–727
- preconventional morality, 366
- predicate calculus, 135
- predicates, 43
 attributes, 43
 binary, 43
- predictions
 from concepts, 178
 from similarity assessment, 155, 155–156
- prediction games, 574
- prediction markets, behavioral finance in, 766–767
- preferences
 for compatibility, in decision making, 309
 in consumer behavior, 758–759
 in cross-modal mapping, 588
 in decision making, 316
 inconsistent, in reason-based choice, 308
 marketing and, 758–759
 in mental accounting, 312
 for music, 778
 in rational theory of choice, 17, 302
 in risky decisions, 304
 in strategy-motivated thinking, 401, 403
- prefrontal cortex, of brain
 aging and, theories for, 652
 anatomy of, 84
 cognitive control in, 84
 for complex cognition, 94–95
 DLPFC, 355
 DMPFC, 355
 emotion in, 369
 episodic memory retrieval and, 85
 FTLD, 72–73, 86
- functional decomposition of, in reasoning tasks, 242–243
- in guided activation model, 679–680
- insight activation in, 477
- mental number line in, 591

- relational reasoning in, 241, 242
 RLPFC, 85
 role in intelligence for, 84–85
 schizophrenia and, 679–680, 680,
 686–688
 socioemotional tasks and, 85
 prefrontal theories, 652
 premature conceptualization, 469
 prepared mind theory, 466
 prepotency, schizophrenia and, 681
 Present Illness Program (PIP), 737–738
 prevention-focused individuals, 404–405
 primary progressive aphasia (PPA), 80
 DCM and, 80
 deductive reasoning and, 78–79
 PPA and, 80
 resolution with, 78
 primates
 brains size of, 530. *See also* brain size
 in cultural intelligence hypothesis,
 534–535
 priming, 314
Principia Mathematica, 492
 private body consciousness, 368
 private information games, 352–353
 acquire-a-company problem in, 352
 communication restriction in, 360
 winner's curse in, 353
 probabilistic reasoning, 144–145
 equiprobability principle in, 144–145
 extensional, 144
 intensional, 144
 subset principle in, 145
 probability theory, 16
 Bayesian approach to, 16
 Bayesian inferences and, 16, 22, 23
 for Bayes theorem, for argumentation,
 293
 beliefs, 12, 16
 circular argument and, 290
 in circular arguments, 290
 classic theory for, 144
 conjunction fallacy and, 337
 degrees of belief and, 12, 16
 explanations, 268–269
 explanations and, quality of,
 268–269
 in heuristics and biases approach to
 judgment, 326
 in negative model, for heuristics and
 biases approach to judgment, 335
 in positive model, for heuristics and
 biases approach to judgment, 329
 in prospect theory, 303
 in rational models, 323
 subjective interpretation of, 16
 support theory and, 338–339
 unification and, from explanations,
 269
 problems, 414
 well-defined *vs.* ill-defined, 462
 problem-based learning (PBL), 748–749
 problem isomorphs, 420–421
 in analogic problem solving, 421
 problem solving, 2
 in academic disciplines, 424–428
 with aging, 656–657
 algorithms in, 419
 analogic, 421
 with analogy, 238–239
 birds-and-trains problem, 415
 components of, 414
 creativity and, 461, 462–463, 471
 data-driven, in medical reasoning,
 740–741
 definition of, 413
 diagrams in, 425, 426–427
 emotion-focused, for aging, 657
 equations in, 425–426
 by experts, 421–423
 forward-driven, in medical reasoning,
 740
 functional fixedness in, 417
 future research applications for,
 428–429
 Gestalt psychology and, 3, 416–418
 with GPS, 419
 heuristics in, 419
 hypothesis-driven, in medical
 reasoning, 740, 740
 idea roadmaps theory and, 467
 initial state of, 414
 insight and, 462–463, 475–478
 interactions within, 415
 isomorphs in, 420–421
 knowledge effects in, 424–428
 mathematical, 417–418
 means-end analysis with, 419, 420
 in medical reasoning, 737, 740–741
 multiconstraint theory and, 239
 neurological model for, 424
 Newell and, 418–420
 nine-dot problem, 417
 operators in, 414
 opportunistic assimilation theory and,
 466
 perception effects in, 424–428
 positive feedback in, aging and, 657
 for practical problems, 417
 through problem spaces, 418–419
 progress monitoring theory and, 424
 for radiation problem, 414
 representational change theory and,
 423–424
 representations in, 414, 416–418, 421
 scientific thinking and reasoning as,
 704, 704–705
 searching in, 414
 Simon and, 418–420
 solution searching in, 420–423
 subjective nature of, 413–414
 surface-based categorization
 in, 423
 with think-aloud protocols, 419
 Tower of Hanoi problem, 415
 for two ropes problem, 417
 well-defined *vs.* ill-defined problems
 in, 462
 problem spaces, 418–419
 in scientific thinking and reasoning,
 705
 in Tower of Hanoi problem, 418
 procedural argumentation, 278
 procedural memory, for music
 performance, 780
 in performance expression,
 780–781
 processing capacity, for causal power
 theory, 219
 process metrics, for creativity, 467
 process models, of higher cognition
 connectionist neural-network, 55–60
 future applications of, 64
 symbolic, 52–55
 symbolic-connectionist, 60–64
 types of, 52
 production systems, 53
Productive Thinking (Wertheimer), 703
 progressive alignment strategy, 251
 progress monitoring theory, 424
 for insight problems, 481
 promotion-focused individuals,
 404–405
 proof-theoretic consistency, 14
 proportional analogy, 236
 propositions
 in similarity assessment, 164
 in structured representations, 43
 prosocial tendencies, among humans,
 533, 534, 535
 prosodic level, of gesture, 641
 prosopagnosia, 210
 prospect theory, 302–303
 EU in, 302–303
 foundation for, 303
 in heuristics and biases approach to
 judgment, 327
 probability in, 303
 risk aversion in, 302–303
 risk seeking in, 303
 prototype extraction tasks, 185–186
 multiple memory systems under, 186
 prototypes
 categorization for, 183–184
 distortions of, 183–184
 exemplar models compared to, 184, 184
 fuzzy set theory for, 182
 in semantics, 182
 psycholinguistics, concepts and, 202
 psychological game (PG) framework,
 346, 348, 358–359
 applications for, 348
 definition of, 358
 emotions in, 358–359
 guilt in, 346, 358
 reciprocity in, 359
 shame in, 358–359
 psychological probability, 329
 psychology. *See* cognitive neuropsychology;

cognitive psychology; correlational psychology, genius study within; developmental psychology; evolutionary psychology; Gestalt psychology; macropsychology; moral psychology; scientific psychology, genius study within psychometaphysics, 178, 195, 202 psychometric tradition, 236 in tripartite theory of rationality, 441–442 psychophysics principles, 326 purity/sanctity domain, in moral judgment, 365

Q

quantifiers, for numbers, 558 quantitative thinking arithmetic and, 592–593 development of, 586 function of, 585 quantitative (*cont.*) in infancy, 586–593 magnitude processing in, 589–592, 600, 601 mental number line in, 589 non-numerical cues for, 587–588 non-symbolic, 585, 586–594 non-symbolic numerical processing, 593–594 number representation, in infants, 587–589, 588–589 numerical discrimination in, 586–587, 586–587 numerical quantities in, 587–588 for numerical symbols, 594–601, 601–602 operational momentum in, 593 representation in, 585–586 symbolic, 585, 594–601 quantity insensitivity, 383 questions, explanations and, 266 Quételet, Adolphe, 492

R

radiation problem as insight problem, 479–480 problem solving for, 414 ratio bias, 340 rational argument, procedures for, 293. *See also* argumentation rationalist theory, of moral judgment, 366–367 conventional morality in, 366 critiques of, 367 as culturally-biased, 367 in developmental psychology, 366 Heinz dilemma under, 366 postconventional morality in, 366 preconventional morality in, 366 rationality, 434–435 abstract selection task for, 438 arationality, 434

Bayes theorem and, 438 competence/performance distinction for, 448 definition of, 434, 449 degrees of, 435 dual process theory for, 438–441 epistemic, 434–435 evolutionary effects on, 436–437 evolutionary influence on, among humans, 436, 436–437, 437 expected value in, 449 Great Rationality Debate, 433, 435–436 individual differences in, 437–438, 448–449 instrumental, 435 intelligence and, 438 irrationality, 434 natural selection and, 436 normative theories of reasoning and, 12 requirements for, 446 tripartite theory of, 441 utility in, 435 rational models Bayes rule, 324 for categories, 184 in economic theory, 323 inference in, 324 for judgment, 323–324 probability coherence in, 323 rational process models, for Bayesian inferences, 25 rational theory of choice, 12, 16–18, 18, 301–302 consistency in, 302 continuity condition in, 17 criticism of, 302 descriptive standards, 302 EU in, 17, 301–302 game theory and, 18 microeconomics and, 18 as normative standard, 302 preferences in, 17, 302 utilities in, 17 value system of, 12 Raven, John C., 236 Raven's Progressive Matrices (RPM) test, 84, 236 aging mind and, 657 reason-based choice, 308–309 information delay and, 308–309 preference inconsistencies in, 308 self-awareness of influences on, 308 reasoning. *See also* business, reasoning in; deductive reasoning; explicit reasoning; legal reasoning; medical reasoning; normative theories, of reasoning; probabilistic reasoning; relational reasoning; role-based relational reasoning; scientific thinking and reasoning abductive inferences and, 270–271 academic history of, 3–4 accuracy, in outcome-motivated thinking, 398–399 aging and, 657–658 analogical, in structured relational representations, 45–46 Bayesian inferences in, 27–28 case-based, 238 categorical, 619 categorization and, 189 dedicated neural systems for, 79 developmental thinking in children and, 520–521 explanation and, 261, 270–271 extensional, 144 folk biology and, 199 inferences in, 2 intensional, 144 moral, 365 motivated thinking and, 404 normative theories of, 11–12 psychology of, development of, 3, 129 rule-based, spatial tasks and, 612 sentential, 137 suppositional, 144 in visuospatial thinking, 619 recall effects on directional outcomes, 394–395 on nondirectional outcomes, 396–397 on strategy-motivated thinking, 402–403 reciprocity, in PG game model, 359 reckoning, 7 recognition. *See also* recognition of objects, in visuospatial thinking analogy, 423 contemporary, 494 facial, brain domain for, 199–200 for false memories, 250 pattern, 422 place and landscape, brain domain for, 200 posthumous, 494 recognition of objects, in visuospatial thinking, 608 models of, 608 structural description models for, 608 viewpoint dependent models for, 608 viewpoint independent models for, 608 recognition paradigm, for false memories, 250 reduced resources theory, 651 attentional resources in, 651 regression artifacts, 329–330 regulators, 632 regulatory focus, in strategy-motivated thinking, 400, 400 regulatory mode, in strategy-motivated thinking, 400, 400 relational deductive reasoning, 144 relational displays, 620 relational language, in mapping, 247 relational models theory, 577 predictive behavior in, 577

- relational priming, in analogical reasoning, 245–246
- semantic alignment in, 245–246
- relational reasoning, 144
- algorithmic model development for, 253–254
 - analogy and, 239–243
 - causal models for, 239
 - computational goals of, 239–240
 - limitations of, 253
 - multiconstraint theory and, 239–240
 - neural implementation of, 254
 - neural substrate of, 241–242
 - in prefrontal cortex, 241, 242
 - representation in, 240
 - schemas and, 252–253
 - semantic similarity and, 240
 - symbolic-connectionist models and, 253
 - translational research for, 254–255
- relational reinterpretation hypothesis
- for animal behavior, 536, 537
 - for human mind, 536–537, 537
- relational shift, 248
- relational thinking, 54
- relationship regulation theory, 577
- predictive behavior in, 577
- reminiscence, 459
- forgetting fixation theory and, 459–461
 - output interference in, 461
- remote association theory, 465–466
- attentional defocusing in, 466
- repeated decisions, 311–312
- as unique, 312
- representations. *See also* analog representations; featured representations; knowledge representation; mental representations; spatial representations; structured representations
- allocentric, 614
 - analogical inference and, 250
 - in analogical reasoning, 253–254
 - analog mental, 39
 - in connectionist neural-network models, 56
 - egocentric, 614, 615
 - external, in visual perception, 425
 - in Gestalt psychology, for problem solving, 421
 - internal, 639–640
 - internal, in visual perception, 424–425
 - for navigation, 613–615
 - of numbers, by infants, 587–589, 588–589
 - for orientation, 614–615
 - in problem solving, 414, 416–418, 421
 - in quantitative thinking, 585–586
 - in relational reasoning, 240
- in symbolic process models, 52–53
- in visual imagery, nature of, 609
- representational change theory
- for insight problems, 423, 481
 - problem solving and, 423–424
- representation and algorithm, 4
- representation of culture, across time, 572–573
- through bias, 573
 - through imitation, 572
 - knowledge mediation in, 572
 - in mini-societies, 572–573
 - through observation, 572
 - scaffolding in, 572
- representativeness heuristic, 331–332
- basic set-inclusion in, 331–332
 - likelihood outcomes in, 331
- represented world, 37
- representing world, 37
- Rescorla-Wagner (RW) model, 214
- backward blocking in, 215
 - modifications to, 215
- retirement funds, behavioral finance for, 765–766
- reverse engineering, Bayesian references and, 25, 26–27
- rhythms, 775
- Richerson, P.J., 533
- Riggs v. Palmer*, 724
- right lobe prefrontal cortex (RLPFC), 85
- risk aversion, 302–303
- riskless decision making, 305–306
- loss aversion and, 305–306
 - semantic framing in, 306
 - status quo and, 305–306
- risk seeking, in prospect theory, 303
- risky decisions, framing of, 303–305
- attitudes towards, 304
 - economic environmental influences on, 304–305
 - global perspectives on, 311
 - local perspectives on, 311
 - majority preferences in, 304
 - with monetary outcomes, 303–304
- RLPFC. *See* right lobe prefrontal cortex
- roadmaps theory. *See* idea roadmaps theory
- “The Road Not Taken” (Frost), 238
- Roe v Wade*, 727
- role-based relational reasoning
- analogy in, 235
 - core properties of, 235
 - in human mind, 536–537, 537
- Rosch, Eleanor, 4, 180
- Rousseau, Jean-Jacques, 494
- RPM test. *See* Raven’s Progressive Matrices test
- rule-based reasoning, spatial tasks and, 612
- rules, for legal reasoning, 722–725
- under English common law, 723–724
 - for hard cases, 722–724
 - multiple and inconsistent, 723–724
- unclear, 722–723
- with wrong answers, 724
- rules of voice-leading, 783
- RW model. *See* Rescorla-Wagner model

S

- sacred and protected values
- cultural context for, 382
 - heuristics for, 383
 - moral judgments and, 382–384
 - quantity insensitivity and, 383
 - rigidity of, 383–384
 - in Value Pluralism model, 382
- sadness, decision making and, 314
- Sapir, Eric, 544
- satisficing, 325
- savants, 495
- scaffolding, 572
- schemas
- in analogical reasoning, 251–253
 - assignment of, 252–253
 - comparison in, 251, 251
 - convergence, 251–252
 - induction factors, 251–252
 - preexisting, 252
 - progressive alignment strategy for, 251
 - relational reasoning and, 252–253
 - transfer of, 252–253
- schizophrenia
- academic literature on, 675
 - capacity allocation and, 684
 - chromosomal influences on, 688
 - cognitive control in, 681
 - cognitive model integration for, 689–690
 - context information sensitivity in, 680, 680, 681
 - delusions in, 675
 - endophenotype for, 685, 685, 689–690
 - environmental factors for, 685
 - epidemiology of, 685
 - episodic memory deficits and, 688–689
 - fetal hypoxia and, 685
 - genetic factors for, 688
 - guided activation model for, 679–681
 - hippocampal volume and, 688–689
 - intermediate phenotypes in, 688
 - long-term memory and, 686, 689
 - neural system abnormalities in, 685–689
 - neuroimaging for, 686
 - pathological mechanisms for, 687–688
 - prefrontal cortex and, 679–680, 680, 686–688
 - prepotency and, 681
 - prevalence of, 685
 - psychopathological difficulties with, 674
- schizophrenia (*cont*)
- social cognitive neuroscience and, 690

- temporal lobe abnormalities and, 688–689
- transmission pattern of, 685
- working memory and, 686, 686–688
- Schwarz, Norbert, 330
- science education, 712–714
- as collaborative group activity, 713
 - constructivist, 713
 - discovery learning approach in, 713
 - inquiry learning approach in, 713–714
 - medical reasoning and, 748–749
 - PBL in, 748–749
 - through self-directed discovery, 714
 - social mechanisms for, 713
 - science knowledge, in medical reasoning, 742, 745
 - Sciences of the Artificial* (Simon), 756
 - scientific psychology, genius study within, 504–505
 - scientific thinking and reasoning
 - abductive reasoning in, 707
 - analogy in, 709, 708–709
 - categorical induction in, 707
 - causal thinking in, 706–707
 - in children, 710–712
 - computational approaches to, 712
 - computers for, 712
 - conceptual change in, 709–710
 - content in, 701
 - data mining in, 712
 - deductive reasoning in, 707–708
 - developmental literature on, for children, 711
 - discovery processes in, 702–703, 703 as hypotheses testing, 703–704, 705–706
 - induction in, 707
 - knowledge elements in, 710
 - literature on, 704
 - parameters of, 701
 - as problem solving, 704, 704–705
 - problem spaces in, 705
 - reasoning processes in, 701, 702
 - research history for, 702–706
 - science education and, 712–714
 - select and test model, in medical reasoning, 739
 - selected explanation, 262
 - complete compared to, 262
 - selection effect, in legal reasoning, 724–725
 - selection tasks, in dual-process theories, 119
 - for explicit reasoning, 120
 - selective attention, in formal thought disorder, 682–683
 - selective modification model, 193
 - selective optimization with compensation (SOC) theory, 652–653
 - expertise in, 664
 - motivation reduction in, 652–653
 - self-adaptors, 632
- self-directed discovery, in science education, 714
- self-explanation, 267, 268
- semantics, 182–183
- category representations in, 182–183
 - fragmentation, within semantic memory, 181–183
 - goals of, 182
 - pathways in, addition of, 182
 - prototypes in, 182
 - standard approach to, 182
- semantic alignment, 245–246
- semantic framing, in riskless decision making, 306
- concept activation through, 306
- semantic memory, 201
- aging and, 660
 - cognitive economy in, 179
 - concepts and, 179–182
 - as development process, 180–181
 - in formal thought disorder, 682
 - fragmentation within, between semantics and memory, 181–183
 - hierarchies within, 179
 - organization models for, 179
- semantic networks
- marker passing in, 44
 - in memory processing, 45
 - mental representations and, meaning from, 38
 - relationship assessment in, 44
 - spreading activation process in, 44–45
 - in structured representations, 44–45
 - in symbolic process models, for mental representations, 55
- sentential reasoning, 137
- sentimental rules theory, 369
- separate attribute evaluation, in decision making, 310–311
- sequential theory, 616–617
- serotonergic genes, attention and, 100
- set-shifting, 659
- SEU. *See* Subjected expected utility
- Shakespeare, William, 492, 494
- shame, in PG model, 358–359
- Shockley, William, 497
- SIAM model, 166
 - direct tests of, 166
 - nodes in, 166
- side-effect effect, 381–382
- Sidis, William James, 495
- Siegel, M., 704
- signal detection model, 325
- signature conventional pattern, 376
 - critique of, 376–377
- signature moral pattern, 375
 - assessment paradigm for, 376
 - critique of, 376–377
- similarity, assessment of, 155, 155
 - abduction reasoning and, 163
 - in alignment-based models, 155, 165–167
 - benefits of, 156
- calculation for, by domain, 172
- categorization in, 170
- among children, 171
- in consumer behavior research, 759
- in contrast models, 161–163
- in developmental psychology, 169–170
- diagnosticity effect and, 170
- as diagnostic tool, 157
- dimensional determination for, 158
- extension effect and, 170
- in featural models, 155, 160–161
- feature learning and, 187
- flexibility of, 157, 169–171
- future applications for, 169–172
- generalizations and, 156
- generic, 171
- in geometric models, 155, 157–160
- inferences from, 156
- judgments on, 170
- in marketing, 759
- MDS in, 157–160
- in perceptual functions, 157
- prediction assessment from, 155, 155–156
- propositions in, 164
- reasonable expectations for, 155
- in relational reasoning, 240
- studies on, 42
- in transformational models, 155, 167–169
- utility of, in cognitive sciences, 156–157, 170–171, 171–172
- similarity of cases, medical reasoning and, 741–742
- psychology of categorization and, 741–742
- similarity tests, for language, 563
- Simon, Herbert, 3, 3, 324–325, 418, 418, 712, 756, 756, 756
- bounded rationality for, 324, 324
- on heuristics, for genius assessment, 500
- problem solving for, 418–420, 704
- problem spaces for, 418–419
- satisficing for, 325
- simultaneity, 651
- single and multiunit recording, in neuroimaging, 74–75
- single nucleotide polymorphisms (SNPs), 92
 - cultural change, 580
- single photo emission computed tomography (SPECT), 77
- situated cognition, 48
 - anti-representationalism in, 48–49
- sketching, 470
- Skinner, B.F., 666
- slippery slope argument (SSA), 282
 - as conditional, 288–289
- Slovic, Paul, 339
- small ideas, in knowledge transfer, 791–793, 793

- vertical transfer of, 793
SMPY. *See* Study of Mathematically Precocious Youth
SNARC effect. *See* spatial-numerical association of response codes effect
SNPs. *See* single nucleotide polymorphisms
social brain hypothesis, 530
 neocortex size as influence on, 531
social cognition, cognitive miser model for, 337–338
social cognitive neuroscience, 690
social collaborative technology, 575–576
 iTunes, 576
 Wikipedia, 575–576, 580
social complexity of species, brain size and, 531
social institutionist model, for moral judgment, 367–368
 cognitive-affective theory in, 368
 components of, 367
 critiques of, 368
 eliciting situations in, 367
 group norms in, 368
 private body consciousness in, 368
social knowledge, in dual-process theories, 115
social psychology, dual-process theories for, 124
social wisdom, 665
society, human
 definition of, 570
 language development as response to, 577
 thinking and, 570
socioemotional selectivity theory (SST), 653, 661
 emotion goal-setting in, 653
 time assessment in, for aging, 655–656
SOC theory. *See* selective optimization with compensation theory
sometimes-opponent-process model, 215
sortalism, 197–198
 categorization and, 197
 in developmental psychology, 197
 essentialism and, 197–198
 object identity in, 198
spatial frames of reference, language for, 554–555
 cross-linguistic differences for, 554–555
 flexibility in, 555
spatial functional neuroimaging, 77–80
 fMRI, 73
 PET, 77–78
 temporal functional neuroimaging and, 85
spatial layout knowledge, 616–617
 individual learning for, 617
 sequential theory for, 616–617
spatial-numerical association of response codes (SNARC) effect, 618
spatial orientation, language for, 558–561, 563
 animal models for, 558–559
 developmental studies, 560–561
 lexical resources for, 559–560
 mental representations of, 559
 speech-shadowing for, 559
spatial relationships, language for, 552–553
 cross-linguistic differences, 552–553
 in developmental psychology, 563
 language-on-language interpretation for, 553, 552–553
 modularity in, 559
spatial representations, 39–42
 definition of, 39–41
 dimensions of, 40
 limitations of, 41
 MDS for, 40–41
 for navigation through environmental space, 613–615
 for orientation to environmental space, 614–615
 strengths of, 41
 triangle inequality in, 41–42
spatial tasks, visual imagery in, 611–613
 augmentation of, through rule-based reasoning, 612
 implicit knowledge for, 612–613
 task decomposition for, 611–612
Spearman, Charles, 3, 236
SPECT. *See* single photo emission computed tomography
speech
 gesture translated into, 636–637
 morpho-phonological encoding in, 678
 normal production model, 676–678
 phonetic encoding in, 678
 physical articulation process for, 678
 roots of gesture in, 640–642
speech abnormalities, 675
 classes of, 675
 idea disassociation as, 675, 675–676
speech disorders
 aphasia, 678
 capacity allocation and, 684
 communication deviance, 690–691
 lemma retrieval and, 678
 morpho-phonological encoding in, 678
 phonetic encoding in, 678
 physical articulation process for, 678
 thought disorders and, 678
speech perception, featured representations for, 42
speech-shadowing, 559
spreading activation, 44–45
 nodes in, 44–45
SSA. *See* slippery slope argument
SST. *See* socioemotional selectivity theory
stare decisis, in legal reasoning, 727
statistics, in Bayesian inferences, integration of, 28
status quo
 bias in, 306
 decisional conflicts from, 307
 from loss aversion, 305–306
 medical outcomes and, 306
 monetary outcomes and, 305–306
Stich, S., 704
stock markets, behavioral finance in, 766
stored solution plans, expertise and, 422
story model, for juries, 730
Strack, Fritz, 333
strategy-motivated thinking, 399–403
 alternative hypotheses for, 400–401
 counterfactual effects on, 401–402
 information processing in, 402, 402–403
 locomotion concerns in, 401
 preferences in, 401, 403
 preferred strategies in, 399–400
 recall effects on, 402–403
 regulatory fit in, 403
 regulatory focus in, 400, 400
 regulatory mode in, 400, 400
Stravinsky, Igor, 497
strength-weight theory, 334
strictly dominant strategy, in game theory, 349
structural description models, 608
structural mapping theory, 465
structural neuroimaging, 73–74
 DTI, 74
 fMRI, 73, 78–80
 MRI, 74
 VBM, 74
structured imagination theory, 457, 465
structured relational representations, 45–46
 alignment in, 46
 analogical reasoning in, 45–46
 analogy in, 45–46, 45–46
 elements of, 45
structured representations, 43–44
 predicates for, 43
 propositions in, 43
 relational, 45–46
 semantic networks in, 44–45
Study of Mathematically Precocious Youth (SMPY), 505
subjected expected utility (SEU), 323
subset principle, 145
 in evolutionary psychology, 145
substances, language for, 550–552
 categorization of, 551
 classificatory tasks for, 550–551
 grammatical determinants for, 550
 grouping of, 551
 language-on-language interpretation for, 552
substances, language for (cont)
 linguistic stimulus of, 552

ontology for, 550
subsumption, 267–268
successful aging, 665–666
support systems, in management science, 763–764
cognitive, 763–764
support theory, 338–339
development history of, 338
intuitive judgment and, 338
suppositional reasoning, 144
Supreme Court, US., 733
symbols
in Bayesian inferences, integration of, 28
inductive inferences and, 30
knowledge components as, 803
in mental representations, 39, 39
symbolic arithmetic, 599–600
analog presentations in, 599–600
linear presentations in, 599
symbolic connectionism, 240
symbolic-connectionist models, 60–64
complexity of relational assumptions in, 63–64
disadvantages of, 63–64
executive functions in, 63
future applications of, 64
LISA, 61–62, 82–84, 241, 241
relational reasoning and, 253
strengths of, 63
vector addition, 61–63
vector multiplication, 60–61
symbolic languages, knowledge transfer for, 789
acquisition of, 795
algebra as, 798, 803, 803
computational modeling for, 795
learning to think and, 790
NLM learning systems for, 795
symbolic process models, 52–55
disadvantages of, 54–55
future applications for, 64
graph matching in, 53
nodes in, 53
notations in, 52–53
physical implementation issues in, 55
processes in, 53–54
production systems in, 53
in relational thinking, 54
representations in, 52–53
semantic content in, for mental representations, 55
strengths of, 54
symbolic quantitative thinking, 585, 594–601
future research applications for, 602
numbers and, cognition for, 585
numerical symbols in, 594–601
processing in, 586
symmetry assumption, 160
syntactic level, of gesture, 641
syntax theory, 373
systematicity principle, 46

T
tacit knowledge, in visual imagery, 609
TASS. *See* The Autonomous Set of Systems
teaching, among animals, 533–534
through observable behavior, 534
technology, culture and, 575–576
effects of/with, 749–750
extended mind hypothesis and, 575
as external representations, 749
for information processing, 750
medical reasoning mediated by, 749–750
for medication management, 750
social collaborative, 575–576
teleological explanations, 264
temporal discounting, 312–313
excessive, 313
fluctuations in, 313
temporal functional imaging, integration of, 85
temporal lobe abnormalities, schizophrenia and, 688–689
temporal logics, 15
Terman, Lewis, 493, 493. *See also* genius
testimony, 292
learning from others through, among children, 523–524
Thaler, Richard, 351
theory change, in children, 521
Theory of Explanatory Coherence, 269
theory of mind, 534–535
for nonhuman animals, 534
theory of relativity
development of, 703
as genius, 494
think-aloud protocols, 419
thinking. *See also* aging, thinking and; decision making; deduction; deductive reasoning; induction; judgments; knowledge transfer; motivated thinking; outcome-motivated thinking; problem solving; quantitative thinking; reasoning; scientific thinking and reasoning
academic history of, 3–4
Bayesian inferences in, 27–28
categorization of, 4
common etymology for, 1–3
as computational, 3
conceptualization in, 1
consciousness and, 2
convergent, 465
culture and, 570
deduction in, 2
definition of, 1
divergent, 465
foresight and, 1
hierarchies within, 1
judgment and, 1
language and, 1, 801–802
logic and, 2
mental representations in, 1–2
psychology of, development of, 3
purpose of, 2
relational, 54
society and, 570
thought. *See also* gesture, in thought
language as, 544–546, 546–547
reorganization of, through language, 548–561
as reproductive, 479
richness of, compared to language, 547
unconscious, 3
thought disorders, 3. *See also*
schizophrenia
case study for, 673–674
definition of, 675–676
dementia, 684–685
diagnoses for, 674
guided activation model for, 679–681
heritable factors for, 674
integrative view of, 674
pathology study for, 674
speech disorders and, 678
threshold public goods game, 360
time
in general metaphors, 556, 556
language for, 556–557
in visuospatial metaphors, 618
time constraints, in causal learning, 227–229
timing, in music performance, 781–782
tip-of-the-tongue (TOT) resolution, 461–462
TMS. *See* transcranial magnetic stimulation
tone, in music, 777
in music composition, 782, 784
tonic-phase hypothesis, 96
TOT resolution. *See* tip-of-the-tongue resolution
Tower of Hanoi problem, 415
gesture and, 632, 643
problem isomorphs in, 420–421
problem spaces in, 418
trade, reluctance to, 305
training, aging and, 664–665
incremented difficulty approach to, 664–665
transactive memory, 571
transcranial magnetic stimulation (TMS), 81–82
concurrent applications with, 82
limitations of, 81–82
transformational models, for similarity assessment, 155, 167–169
in AI, 167
alignment-based models and, 168–169
complexity of transformation in, 168
transmission of culture, across time, 572–573
through bias, 573
through imitation, 572
knowledge mediation in, 572

in mini-societies, 572–573
 through observation, 572
 scaffolding in, 572
 trial instructions, for juries, 729
 triangle inequality, 41–42
 assumption, 160–161
 tripartite theory of rationality, 441
 beliefs in, 443
 individual differences in, 441–445
 intelligence in, 443, 443
 mindware for, 445–446
 psychometrics and, 441–442
 trolley dilemma, 377–379
 DDE and, 378, 379
 directness factor in, 378–379
 double causal contrast theory in, 379–380
 loop idea in, 379
 moral grammar theory and, 377–378
 truth, principle of, 138–140
 in directional outcomes, 393
 fully explicit models for, 139
 illusory inferences and, 142, 142
 implicit models for, 139
 inferences in, 139–140
 tables for, 137–138
 testing for, 139
 Tversky, Amos, 4, 160, 326. *See also* heuristics and biases approach, to judgment
 on conjunction fallacy, 337
 judgment heuristics for, 322–323
 Tweney, R.D., 704, 704
 twin studies, epigenetic influences in, 94
 two ropes problem, 417
 two-sided arguments, 281
 type 1 processes, in dual process theory, 116, 121–124
 autonomy in, 440
 cognitive bias and, 126
 as contextualized, 127–128
 in continuum, 124
 explicit reasoning and, 128–129
 intuitive judgment in, 128, 128–129
 multiple systems in, 125–126
 for rationality, 439–441
 speed of processing in, 128–129
 type 2 process override of, 441
 variety of, 441
 type 2 processes, in dual process theory, 116–117, 121–124
 as abstract, 127–128
 in continuum, 124
 explicit reasoning and, 128–129
 intelligence and, 122
 multiple systems in, 125–126
 as nonautonomous, 440–441
 normative responses, 127
 overriding type 1 processes, 441
 for rationality, 439–441
 speed of processing in, 128–129
 in working memory, 121–122

U

Ullman, Shimon, 167
 uncertainty, decision making under, 302–305
 outcomes from, 302
 prospect theory and, 302–303
 risky decision framing, 303–305
 unclear rules, for legal reasoning, 722–723
 for judges, 723
 for Legal Realists, 723
 unconscious cognition, as below-threshold activation, 464. *See also* implicit cognition, creativity and
 unconscious thought, 3
 unconstrained hypothesis generation, for insight, 478
 unexpected findings, causal thinking and, 706–707
 unification, from explanations, 267–268
 probability and, 269
 universalist perspective, on language, 544
 utilities, in rational choice theory, 17
 EU, 17
 utility conditionals, in fallacies, 288

V

Value Pluralism model, 382
 values. *See* sacred and protected values
 Van Damme, Eric, 359
 VBM. *See* voxel-based morphometry
 vector addition symbolic-connectionist models, 61–63
 conjunctive coding in, 62
 in DORA, 62–63
 in LISA, 61–62
 synchrony of firing in, 61
 vector multiplication symbolic-connectionist models, 60–61
 tensors in, 60–61
 verbal analogical reasoning, 242
 verbal overshadowing, of insight, 478–479
 vertical precedent, in legal reasoning, 726–727
 vertical transfer, in knowledge transfer, 793
 viewpoint dependent models, 608
 viewpoint independent models, 608
 congruency effect in, 608
 VIM. *See* violence inhibition mechanism
 violence inhibition mechanism (VIM), 376
 modular, 376
 vision, navigation and, 613
 vista space, 607
 visual agnosia, 187–188
 visual attention, 339–340
 visual illusion paradigm, 326–327
 visual imagery, 608–612
 attention allocation for, 611
 characteristics of, 608–609
 in complex spatial tasks, 611–613
 mental precepts and, 609

mental rotation strategies for, 610, 610–611

motor, gesture and, 642–643
 nature of representations in, 609
 neural basis for, 609–610
 neuroimaging for, 609
 object interaction and, 610–611
 tacit knowledge in, 609
 transformation of, 610

visual impedance effect

visuospatial metaphors, 618
 gesture and, 642
 SNARC effect, 618
 time in, 618

visuospatial thinking
 abstract, 618–623
 in animals, 606–607
 attention allocation in, 611
 categorical reasoning in, 619
 complex spatial tasks, 611–613
 content in, 619
 core knowledge systems, 607
 in deductive reasoning, 618–620
 environmental space in, 607, 613–618

figural space in, 607
 forms of reasoning in, 619
 functions of, 606
 future research strategy for, 624
 general applications of, 606
 gesture and, 639, 642, 644
 with graphical displays, 620–623
 imagining objects in, 608–611
 individual differences in, 619–620
 intelligence and, 606
 linear reasoning in, 619
 mental precepts and, 609
 mental rotation strategies for, 610, 610–611
 in metaphor, 618
 object interaction and, 610–611
 for objects, 607–613
 recognition of objects in, 608
 tacit knowledge in, 609
 training strategies, 624
 vista space in, 607
 visual imagery in, 608–612
 in working memory theories, 606

vivid information, 766

vocal control, 779

Von Gudden, Bernhard, 67

Vos Savant, Marilyn, 495, 497

voxel-based morphometry (VBM), 74

Vygotsky, Lev, 3

W

warmth ratings

 insight and, 464

 intuition and, 464

Wason, Peter, 3, 118, 703–704

 selection tasks, in dual-process theories, 119

Watson, James, 502

- wayfinding (*cont*)
- wayfinding, 617–618
- by gender, 617
 - with landmarks, 617
- weakly dominant strategy, in game theory, 349–350
- Weber-Fechner law, 587
- mental number line under, 589
 - non-symbolic numerical processing under, 594
 - numerical symbol comparison under, 595
- well-defined problems, 462
- Wertheimer, Max, 3, 416, 703, 703. *See also* Gestalt psychology
- Whorf, Benjamin, 544
- Whorf-Sapir hypothesis, 544–545, 561, 562
- cues in, 545
- Wikipedia, 575–576, 580
- winner's curse, 353
- Wisconsin Card Sorting Task, 659
- wisdom
- from aging, 665–666
 - creativity and, 665, 665
 - definition of, 665
 - genius and, 665
 - social, 665
 - successful aging with, 665–666
- The Wisdom Paradox* (Goldberg), 665
- word substitution, in formal thought disorder, 678
- working memory
- aging and, 660
 - analogical mapping and, 248
 - analogical reasoning and, 240–241
 - attention and, 98
- cognition and, 95
- cognitive neurogenetics and, 94–98
- complex cognition and, 95
- dopamine-system-related genes in, 95–96, 96–97
- dual-process theories and, 121–122
- genetic polymorphisms in, 98
- genotypic effects on, 96–97
- gesture and, 639
- in insight problems, 483
- moral dilemmas and, 380
- non-dopaminergic genes and, 97–98
- schizophrenia and, 686, 686–688
- tonic-phase hypothesis for, 96
- type 2 processing in, 121–122
- visuospatial thinking and, 606
- Wundt, Wilhelm, 679

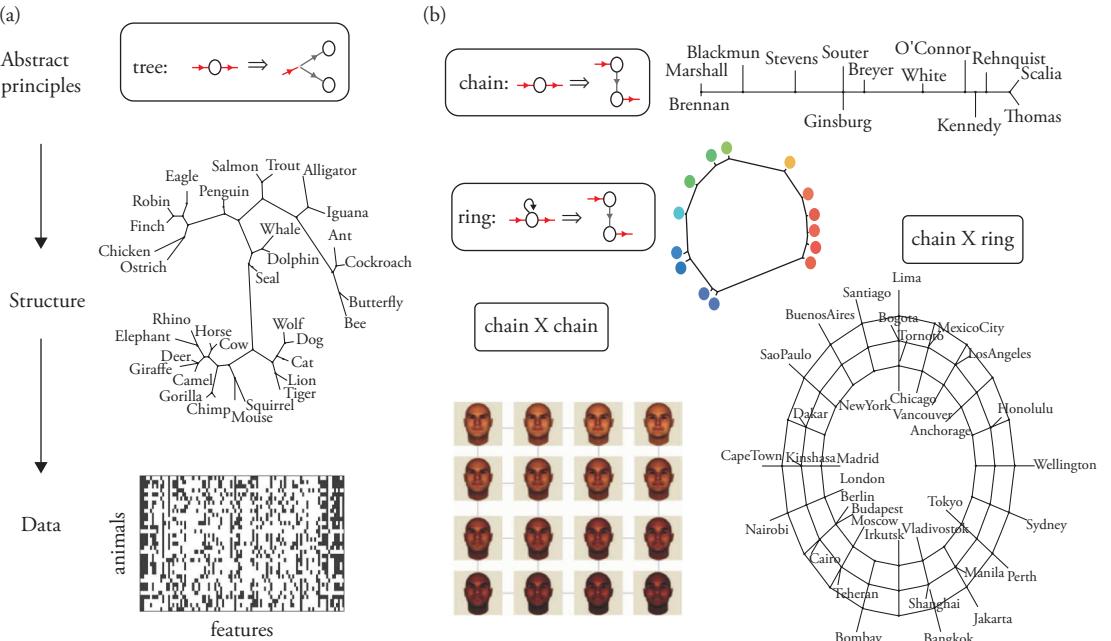


Fig. 3.3 Kemp and Tenenbaum (2008) showed how the form of structure in a domain can be discovered using a hierarchical Bayesian model defined over graph grammars. At the bottom level of the model is a data matrix D of objects and their properties, or similarities between pairs of objects. One level up is a graph describing how properties are distributed over objects. Intuitively, objects nearby in the graph are expected to share properties. At the highest level grammatical rules specify the form of structure in the domain—rules for growing graphs of a constrained form out of an initial seed node. A search algorithm attempts to find the combination of a form grammar F and graph G generated by that grammar which jointly receive highest posterior probability $P(F, G|D)$. (a) Given observations about the features of animals, the algorithm infers that a tree structure best explains the data. The best tree found captures intuitively sensible categories at multiple scales. (b) The same algorithm discovers that the voting patterns of U.S. Supreme Court judges are best explained by a linear “left-right” spectrum, and that subjective similarities among colors are best explained by a circular ring. Given images of faces varying in two dimensions, race and masculinity, the algorithm successfully recovers the underlying two-dimensional grid structure. Given proximities between cities on the globe, the algorithm discovers a cylindrical representation analogous to latitude and longitude.

Multipolymorphic analysis

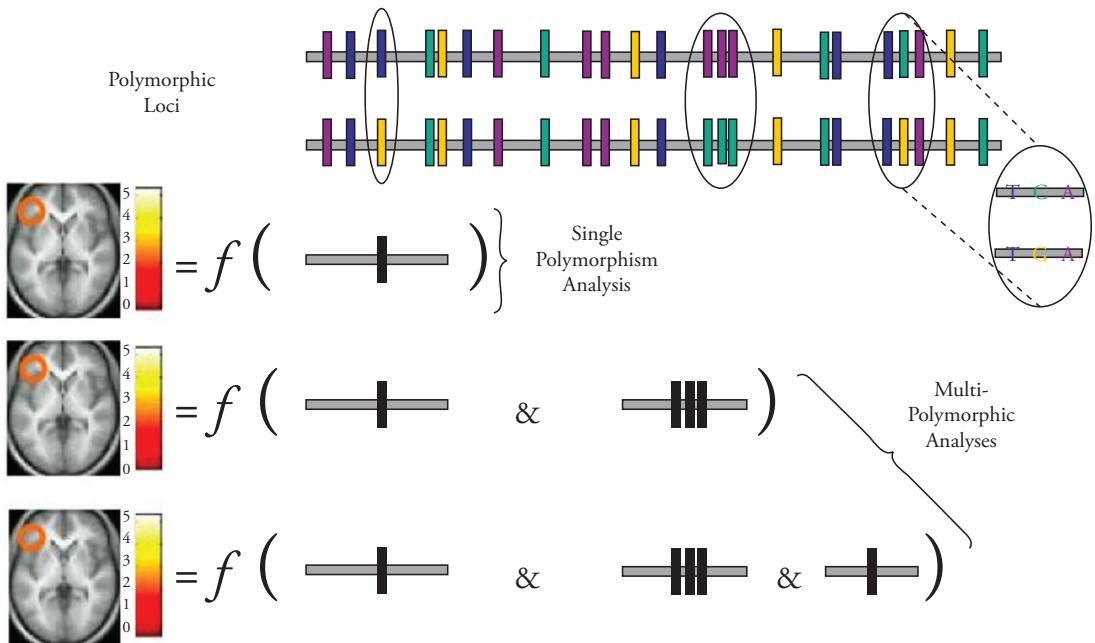


Fig. 7.2 Multipolymorphic analysis. Multiple polymorphic sites in the genome influence a psychological phenotype and its underlying brain activity (brain-based intermediate phenotype). The combined, complexly interactive contributions of these polymorphisms influence neural activity and cognitive function. Most cognitive neurogenetic research to date has focused on variation at individual polymorphisms. These findings are instructive, particularly where only one candidate polymorphism has been implicated by prior findings. However, a growing number of cognitive neurogenetic studies are testing the association of variations at multiple polymorphisms with brain-based intermediate phenotype(s). The figure illustrates a generic multipolymorphic analysis. Polymorphic loci are indicated within two example DNA sequences (where the upper sequence differs from the lower sequence). Brain activity in a region of interest is predicted as a function of one, two, or three of these polymorphisms. Development of multipolymorphic analyses will help construct a better understanding of the complex genetic “networks” that contribute to cognitive brain function and human psychology.

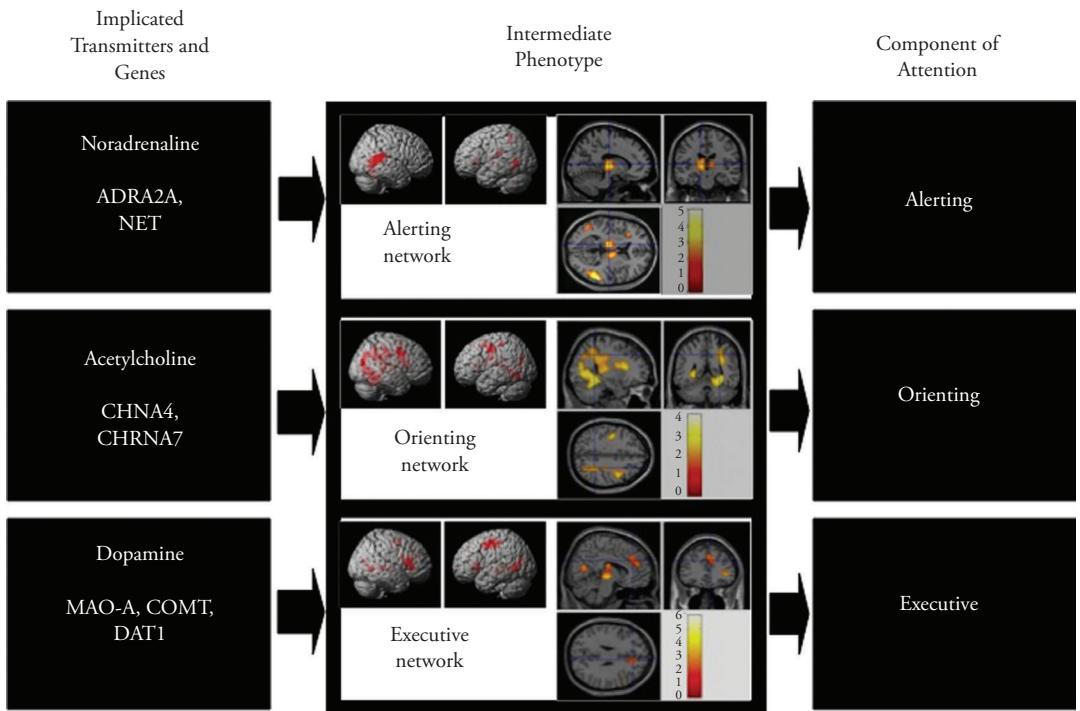


Fig. 7.3 Parsing components of attention in cognitive neurogenetic studies with the Attention Networks Test (ANT). Variations in a set of candidate genes have been correlated with behavioral performance or brain activity for distinct components of attention using the ANT. To generate candidate genes for cognitive neurogenetic studies—in this case for studies on attention—it is important to rely on multiple sources of converging evidence. First, neuroimaging and lesion data pointed to separable neural networks that carry out different aspects of attention. Secondly, pharmacological manipulations demonstrated that noradrenergic modulation can influence the efficiency of the alerting network, while cholinergic modulation and dopaminergic modulation can influence orienting and executive control of attention, respectively. With regard to neural intermediate phenotypes, brain networks were identified where attention-related activity overlaps with patterns of gene expression. It was hypothesized that variation in gene sequence should correlate with individual differences in neural activity associated with components of attention. In the case of executive control of attention, hypotheses of this kind have been supported for brain regions that are targets of dopamine innervation such as the frontal midline, lateral prefrontal areas, and basal ganglia.

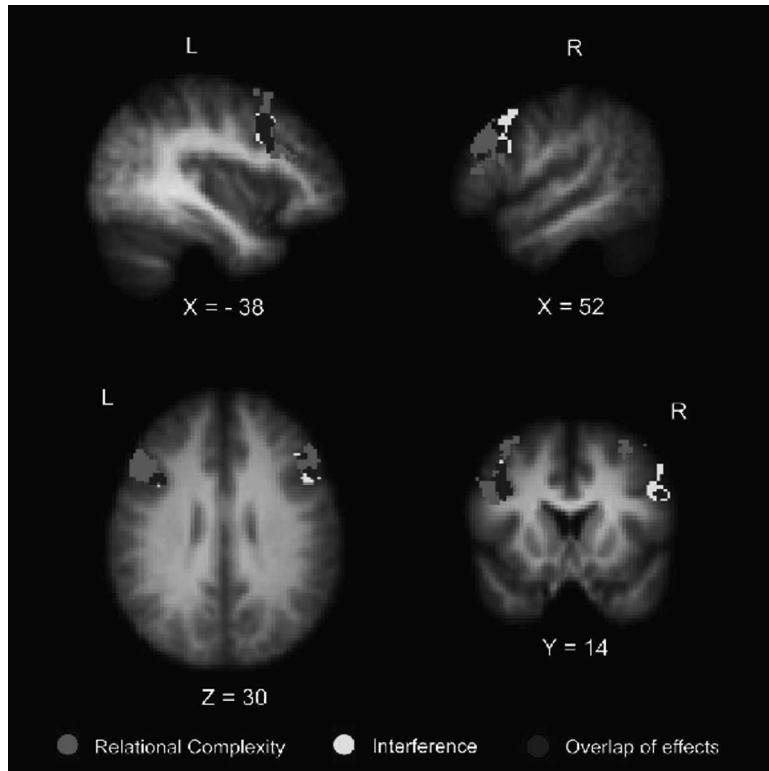


Fig. 13.3 Neuroimaging results from Cho et al. (2010). Regions showing the main effects of relational complexity (shown in red), interference (shown in yellow; small volume corrected, uncorrected cluster-forming threshold $T > 2.3$, corrected cluster extent significance threshold, $p < .05$), and regions where main effects overlapped (blue) within an a priori defined anatomical ROI mask of the bilateral MFG and IFG pars opercularis and pars triangularis. R, right; L, left. Coordinates are in MNI space (mm). (Reprinted by permission.)

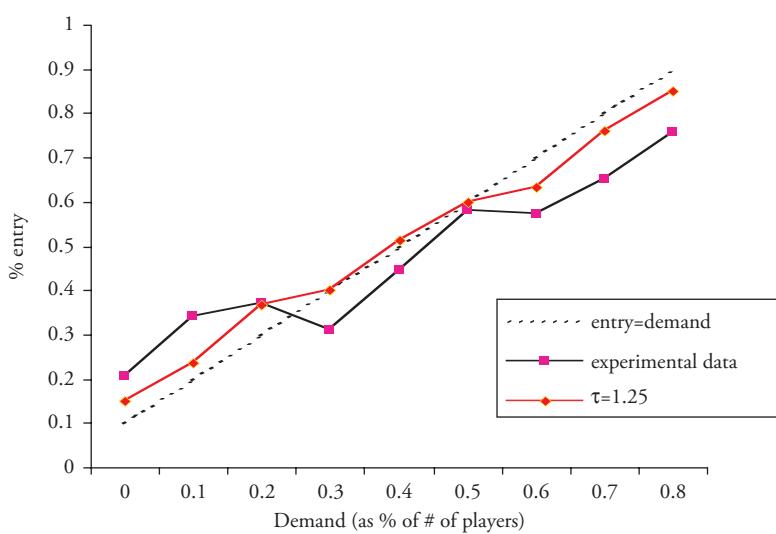


Fig. 18.2 Predicted and observed behavior in entry games.

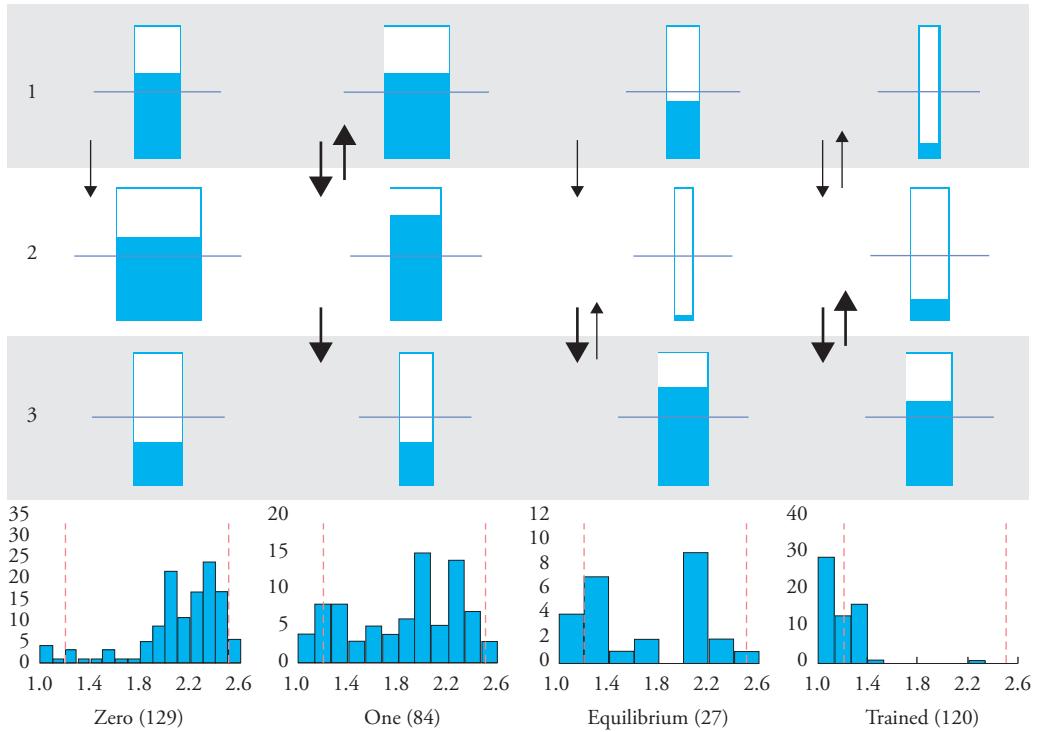


Fig. 18.4 An icon graph of visual attention in three rounds of bargaining (1, 2, and 3) and corresponding distributions of offers. Each column represents a different “type” of person-trial classified by visual attention.

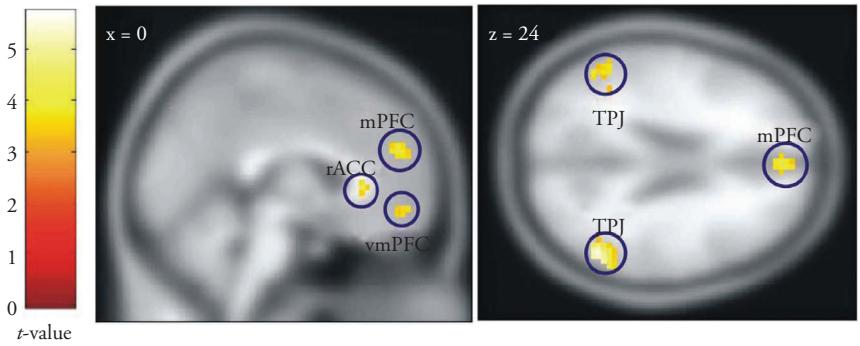


Fig. 18.7 Brain regions more active in level-2 reasoners compared to level-1 reasoners (classified by choices), differentially in playing human compared to computer opponents. mPFC, medial prefrontal cortex; rACC, rostral anterior cingulate cortex; TPJ, temporoparietal junction; vmPFC, ventromedial prefrontal cortex. (From Coricelli & Nagel, 2009, Fig. S2a.)

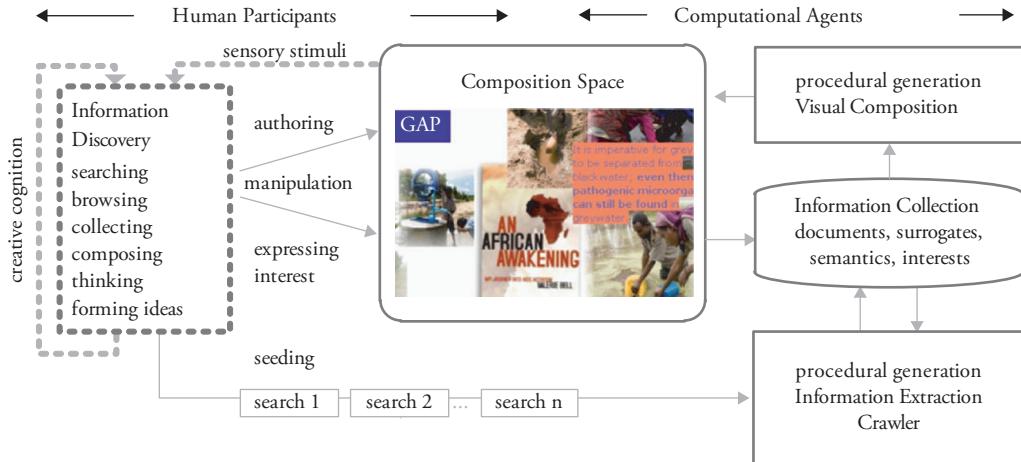


Fig. 23.9 Computational agents support human participants in the combinFormation digital tool for information discovery, which involves browsing through collections returned by search engines and forming navigable compositions of relevant results (Kerne, Koh, Smith, Webb, & Dworaczyk, 2008b). Problems are iteratively reformulated, representations are manipulated, and solutions constructed, often involving integration of multiple information resources (Kerne & Smith, 2004).

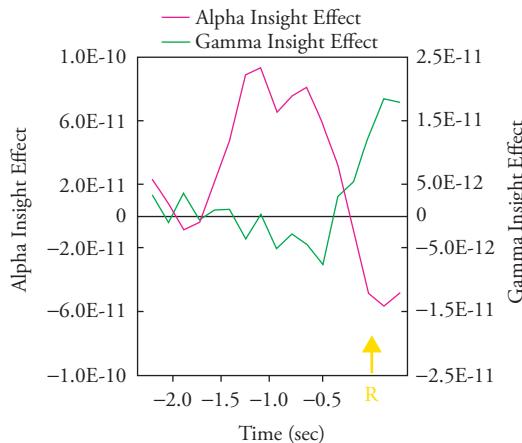


Fig. 24.1 The time course of the insight effect. Electroencephalogram (EEG) alpha power (9.8 Hz at right parietal-occipital electrode) and gamma power (39 Hz at right temporal electrode) for the insight effect (i.e., correct insight solutions minus correct noninsight solutions, in μV^2). The left y-axis shows the magnitude of the alpha insight effect (purple line); the right y-axis applies to the gamma insight effect (green line). The x-axis represents time (in seconds). The yellow arrow and R (at 0.0 s) signify the time of the button-press response indicating that a solution was achieved. Note the transient enhancement of alpha on insight trials (relative to noninsight trials) prior to the gamma burst signifying insight. (Reproduced from open source article by Jung-Beeman et al., 2004.)

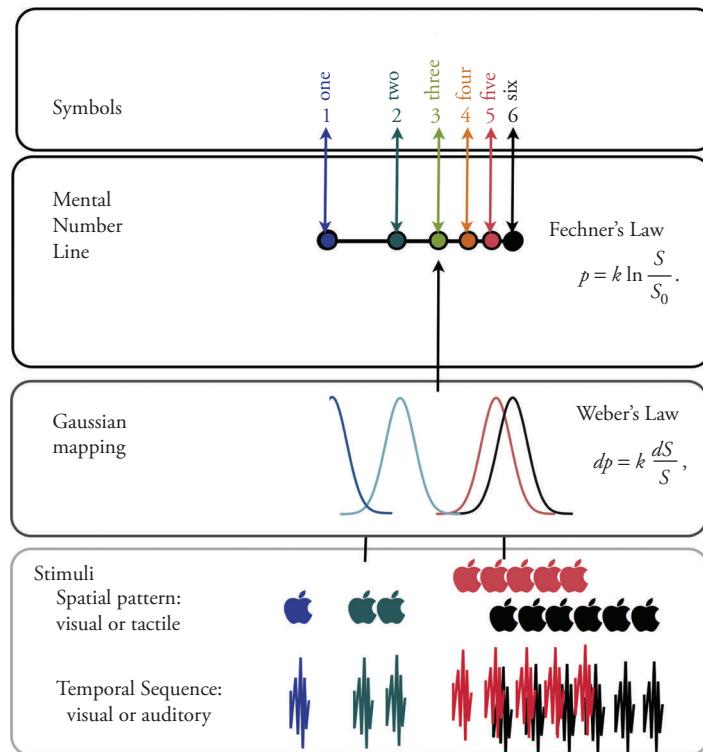


Fig. 30.1 Illustration of a logarithmically-scaled mental number line.