

PHILOSOPHY, MIND, AND COGNITIVE INQUIRY

STUDIES IN COGNITIVE SYSTEMS

James H. Fetzer
University of Minnesota, Duluth
Editor

EDITORIAL BOARD

Fred Dretske
University of Wisconsin, Madison

Ellery Eells
University of Wisconsin, Madison

Alick Elithorn
Royal Free Hospital, London

Jerry Fodor
Rutgers University

Alvin Goldman
University of Arizona

Jaakko Hintikka
Florida State University

Frank Keil
Cornell University

William Rapaport
State University of New York at Buffalo

Barry Richards
University of Edinburgh

Stephen Stich
Rutgers University

Lucia Vaina
Boston University

Terry Winograd
Stanford University

VOLUME 3

PHILOSOPHY, MIND, AND COGNITIVE INQUIRY

Resources for Understanding Mental Processes

Edited by

DAVID J. COLE

Department of Philosophy

University of Minnesota, Duluth, U.S.A.

JAMES H. FETZER

Department of Philosophy

University of Minnesota, Duluth, U.S.A.

and

TERRY L. RANKIN

IBM AI Support Center, Palo Alto

California, U.S.A.



KLUWER ACADEMIC PUBLISHERS

DORDRECHT / BOSTON / LONDON

Library of Congress Cataloging-in-Publication Data



Philosophy, mind, and cognitive inquiry : resources for understanding mental processes / edited by David J. Cole, James H. Fetzer, and Terry L. Rankin.

p. cm. -- (Studies in cognitive systems)

Includes bibliographical references.

ISBN-13: 978-94-010-7340-0

1. Intellect. 2. Human information processing. 3. Cognitive science. I. Cole, David John, 1948-. II. Fetzer, James H., 1940-. III. Rankin, Terry L. IV. Series.

BF431.P4824 1989

128'.2--dc20

89-36760

ISBN-13: 978-94-010-7340-0 e-ISBN-13: 978-94-009-1882-5

DOI: 10.1007/978-94-009-1882-5

Published by Kluwer Academic Publishers,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

Kluwer Academic Publishers incorporates
the publishing programmes of
D. Reidel, Martinus Nijhoff, Dr W. Junk and MTP Press.

Sold and distributed in the U.S.A. and Canada
by Kluwer Academic Publishers,
101 Philip Drive, Norwell, MA 02061, U.S.A.

In all other countries, sold and distributed
by Kluwer Academic Publishers Group,
P.O. Box 322, 3300 AH Dordrecht, The Netherlands.

Printed on acid-free paper

All Rights Reserved

© 1990 by Kluwer Academic Publishers

Softcover reprint of the hardcover 1st edition 1990

No part of the material protected by this copyright notice may be reproduced or
utilized in any form or by any means, electronic or mechanical,
including photocopying, recording, or by any information storage and
retrieval system, without written permission from the copyright owner.

To
Fred Dretske

TABLE OF CONTENTS

SERIES PREFACE	ix
ACKNOWLEDGEMENTS	xi
DAVID J. COLE / Cognitive Inquiry and the Philosophy of Mind	1
PROLOGUE: WHAT IS MIND?	
DANIEL C. DENNETT / Current Issues in the Philosophy of Mind	49
PART I: COMPUTATIONAL CONCEPTIONS	
FRED DRETSKE / Machines and the Mental	75
ZENON W. PYLYSHYN / What's in a Mind?	89
PART II: CONNECTIONIST CONCEPTIONS	
WILLIAM RAMSEY, STEPHEN STICH, and JOSEPH GARAN / Connectionism, Eliminativism, and the Future of Folk Psychology	117
PAUL SMOLENSKY / On the Proper Treatment of Connectionism	145
PART III: REPRESENTATIONAL CONCEPTIONS	
JERRY A. FODOR / Semantics, Wisconsin Style	209
KEN SAYRE / Cognitive Science and the Problem of Semantic Content	229

PART IV: MENTALITY AND INTENTIONALITY

RODERICK M. CHISHOLM / The Primacy of the Intention	255
JOHN R. SEARLE / Intentionality and Its Place in Nature	267

PART V: EPISTEMOLOGY AND COGNITION

HILARY PUTNAM / Why Reason Can't Be Naturalized	283
ALVIN I. GOLDMAN / The Relation Between Epistemology and Psychology	305

PART VI: THE MENTAL AND THE PHYSICAL

PATRICIA KITCHER / Two Versions of the Identity Theory	347
WILLIAM BECHTEL / A Bridge Between Cognitive Science and Neuroscience: The Functional Architecture of Mind	363

EPILOGUE: CONFLICTING CONCEPTIONS

JAMES H. FETZER / Language and Mentality: Computational, Representational, and Dispositional Conceptions	377
SELECTED BIBLIOGRAPHY	403
INDEX OF NAMES	427
INDEX OF SUBJECTS	435

SERIES PREFACE

This series will include monographs and collections of studies devoted to the investigation and exploration of knowledge, information, and data-processing systems of all kinds, no matter whether human, (other) animal, or machine. Its scope is intended to span the full range of interests from classical problems in the philosophy of mind and philosophical psychology through issues in cognitive psychology and sociobiology (concerning the mental capabilities of other species) to ideas related to artificial intelligence and computer science. While primary emphasis will be placed upon theoretical, conceptual, and epistemological aspects of these problems and domains, empirical, experimental, and methodological studies will also appear from time to time.

No problem within the field of cognitive inquiry is more difficult than that of developing an adequate conception of the nature of mind and of its mode of operation. Our purpose in compiling the present volume has been to contribute to the pursuit of this objective by bringing together a representative cross-section of the principal approaches and the primary players who are engaged in contemporary debate on these crucial issues. The book begins with a comprehensive introduction composed by David Cole, the senior editor of this work, which provides a background for understanding the major problems and alternative solutions, and ends with a selected bibliography intended to promote further research. If our efforts assist others in dealing with these issues, they will have been worthwhile.

J.H.F.

ACKNOWLEDGEMENTS

The selection of papers in this volume are reprinted from the following sources:

- Dennett, D.C.: 1978, 'Current Issues in the Philosophy of Mind', *American Philosophical Quarterly* **15**, 249–261.
- Dretske, F.: 1985, 'Machines and the Mental', *Proceedings and Addresses of the American Philosophical Association* **59**, 23–33.
- Pylyshyn, Z.W.: 1987, 'What's in a Mind?', *Synthese* **70**, 97–122.
- Ramsey, W., Stich, S. and Garon, J.: 1989, 'Connectionism, Eliminativism, and the Future of Folk Psychology', in W. Ramsey, Stich, S. and Garon, J. (eds.) (forthcoming).
- Smolensky, Paul: 1988, 'On the Proper Treatment of Connectionism', *Behavioral and Brain Sciences* **11**, 1–74.
- Fodor, J.A.: 1984, 'Semantics, Wisconsin Style', *Synthese* **59**, 231–250.
- Sayre, K.: 1987, 'Cognitive Science and the Problem of Semantic Content', *Synthese* **70**, 247–269.
- Chisholm, R.M.: 1984, 'The Primacy of the Intentional', *Synthese* **61**, 89–109. Revised by the author for this volume.
- Searle, J.R.: 1984, 'Intentionality and Its Place in Nature', *Synthese* **61**, 3–16.
- Putnam, H.: 1982, 'Why Reason Can't Be Naturalized', *Synthese* **52**, 3–23.
- Goldman, A.I.: 1985, 'The Relation Between Epistemology and Psychology', *Synthese* **64**, 29–68.
- Kitcher, P.: 1982, 'Two Versions of the Identity Theory', *Erkenntnis* **17**, 213–228.
- Bechtel, W.: 1983, 'A Bridge Between Cognitive Science and Neuroscience: The Functional Architecture of Mind', *Philosophical Studies* **44**, 319–330.
- Fetzer, J.H.: 1989, 'Language and Mentality: Computational, Representational, and Dispositional Conceptions', *Behaviorism* **17**, 21–39.

DAVID J. COLE

COGNITIVE INQUIRY AND THE PHILOSOPHY OF MIND

Artificial Intelligence cannot avoid philosophy. If a computer program is to behave intelligently in the real world, it must be provided with some kind of framework into which to fit particular facts it is told or discovers. This amounts to at least a fragment of some kind of philosophy, however naive. Here I agree with philosophers who advocate the study of philosophy and claim that one who purports to ignore it is merely condemning himself to a naive philosophy.

— John McCarthy (1988)¹

Artificial intelligence deals with the most concrete and modern of artifacts, digital computers, and at the same time raises abstract and philosophical questions. Can a computer be genuinely intelligent? An answer to this question presupposes that we have an answer to a more general question: What is intelligence? Can a computer simulate intelligence? We then need to know what the difference is between merely simulating intelligence and being intelligent. Can a computer *understand* a language? What *is* a language? Can a computer have a *mind*? Can a sequential digital computer do what a massively parallel neuronal system can? Can a computer have *beliefs, desires, intentions*? Can a computer make *mistakes*? Have *emotions*? Experience *sensations*? Be *creative*? Are computers more like programmed texts or like the authors of such texts?

These are clearly philosophical questions. They are representative of the philosophical issues that arise in cognitive science. As is often the case in a new science, many exciting philosophical and conceptual questions are explored in the search for the most productive approaches and most correct general understanding of the phenomena. Although such interest accompanies the birth of any new science, the subject matter of cognitive science is one of the most intrinsically interesting that exists – the mind.

It is not surprising then that these are exciting times in cognitive science. Workers in cognitive psychology, linguistics, philosophy, artificial intelligence, and neuroscience not only enjoy new developments in their own

areas, but a shared sense that all are working on a common project, the achievement of an understanding of mental phenomena. Underlying this interdisciplinary project are certain philosophical presuppositions about its possibility and character. The papers in this collection represent recent philosophical discussions of various aspects of the cognitive science project, thus conceived.

With the possible exception of neuroscience, there have been profound philosophical changes in each of these disciplines since World War II that have contributed to the emergence of cognitive science. There have, of course, been other changes as well. Although it is obvious that science drives much technological change, it is also the case that technology drives science. Much of good experimental science involves the skilled exploitation of the latest technology. This has been true from the time of the clocks and lenses that propelled the rise of modern science to the laboratories of the present day. Indeed, it would be difficult to imagine what present laboratory science would be in the absence, specifically, of electronic technology.

While it could be an interesting project in philosophy of science to go chronologically through the list of chemistry or physics Nobel laureates from the turn of the century and ask, "What devices made this work possible?", it would not be a very challenging task in the case of cognitive science. There is no doubt that cognitive science owes its present form to a particular technological development, the programmable digital computer. The influence of the computer is on two levels, however, which distinguishes the relation between cognitive science and its enabling technology from the relationship generally characteristic of science. There is the usual relation of enablement, of course. Virtually all work in artificial intelligence, and much in psychology and neuroscience, is either made possible or is greatly facilitated by computers. The computer enables research much as it does in other areas of science, permitting data collection and analysis, and modeling.

But in cognitive science the computer also serves a theoretical and philosophical role. The concept of the computer itself serves as a theoretical model for cognitive science. The most conceptually fruitful distinction has turned out to be that between hardware and software. Although this distinction was implicit from the very beginning in the theoretical notion of an abstract machine, such as a Turing machine, it did not actually come to exert a widespread influence on thought until it became clear that there could be quite distinct disciplines – computer science and computer engineering – dealing with the same phenomenon (and an artifact, yet). If the computer could be fruitfully investigated and described on two levels – one physical,

the other functional – why not the mind itself?

Within philosophy proper, this suggestion assumed a more revolutionary form. Now there arose the prospect that the ordinary language terms used every day to describe our mental lives – the vocabulary of beliefs, desires, emotions, and so forth – could turn out to be a description of the working of the brain at a functional level. Thus the relation between the twin branches of computer science might serve as a model used in understanding and analyzing the mental. This specific influence of the emergence of computer technology is a matter we shall pursue.

Postwar philosophical changes no doubt affected philosophy itself most directly, but they influenced other disciplines dealing with mental phenomena as well. It is difficult to characterize these changes generally. The main change was a challenge to certain presuppositions of empiricism. This resulted in profound changes in psychology, linguistics and philosophy. The most notable of these were the emergence of transformational grammar and a corresponding methodology in linguistics and the decline of behaviorism in psychology. These changes in turn have had many other effects. One of these has been to bring philosophers together with empirical scientists to work on the problems of the mental.

Several considerations have guided the structure of this collection. We want to provide readers with representative work by leading philosophers working on problems in cognition and the nature of mind. Natural divisions within philosophical work in these areas is reflected in the section divisions of the book. These divisions occur on two dimensions: divisions between distinct conceptions of the nature of mind, which are competing fundamental theories of the same general subject; and divisions based on the different specific problem areas. Along the first dimension, we distinguish computational, connectionist, representational, and intentional conceptions of mind and accordingly devote sections to each. Along the second dimension, we provide sections dealing with the relation of epistemology to cognition/psychology; and with the relation of the mental to the physical.

Computational theories of mind hold that a perspicuous account of mental phenomena and intelligent behavior is possible by viewing the mental as a composite of processes that are computational. Computationalism is therefore congenial to computer models of cognitive processes. Computations are performed on the basis of purely formal or physical properties of the states of the system. A computationalist need not hold that any of these computational subprocesses have semantic properties, nor that there is anything in the system which corresponds directly to beliefs, knowledge, or other categories

employed by explanations in ordinary 'folk' psychology. In holding that computational explanations are possible, however, computationalism is committed to a level of description that is at least one remove in abstractness from the underlying physiological processes which actually compute.

Representational views, by contrast, emphasize the semantic properties of at least some mental states. A representationalist may believe that there are internal correlates of belief and mental image, and may appeal to the notion of an internal representation system, i.e., a mental language ('mentalese'). On this account, therefore, representationalism is compatible with computationalism, as long as the representationalist holds that purely formal transformations can preserve semantic properties.

Both computational and representational views were developed during the period that classical von Neumann serial architectures were the paradigm of computer design. This structure fits nicely with a rules and representations model of mentality, where cognition – especially inference – is captured by the application of formal rules to syntactically structured internal representations. Recently, however, connectionist or massively parallel models of distributed information processing have arisen as an alternative to the centralized sequential classical architectures. The models apparently much more closely resemble the neurophysiology of the brain than do classical machines. Connectionist approaches are still computational, but they are not necessarily compatible with representational theories of mind, nor even the everyday attribution of beliefs and desires of ordinary 'folk psychology'. The implications of connectionism for a general account of mind is the subject of much current debate, including the papers in Section II of this collection.

Views of the mental emphasizing intentionality are the most removed from computational accounts. Intentionality is the property of 'aboutness' that is often taken to be characteristic of mental states. Indeed, the 19th century philosopher-psychologist Franz Brentano took intentionality to be the defining mark of mentality, setting mind apart from the rest of nature. Advocates of these views of mind reject the idea that, say, a computer processing sentences as its sole input could have any mental properties precisely because its states are devoid of intentionality: only a human interpreter of these sentences could take them to be about anything.

MAIN PROBLEMS IN COGNITIVE SCIENCE

Cognitive science is interdisciplinary. It is not surprising, therefore, that some of the problems surrounding the mental can be cast as unresolved questions

that exist concerning the relations between various disciplines dealing with the mind.² For example:

1. To what extent should psycholinguistic theories pay heed to formal grammars generated by linguists? On the one hand, both of these disciplines deal with language competence. On the other, cognitive psychology has not found formal logic to be an especially rich source of insight into actual thought processes.
2. How is psychology related to neurophysiology? On the one hand, both disciplines deal with the operation of the central nervous system. On the other, many cognitive psychologists produce theories that make no mention of neurons or their functioning.
3. How is philosophy related to empirical disciplines such as neurophysiology and psychology? On the one hand, philosophy deals with conceptual and *a priori* questions. On the other, psychology until recently was a branch of philosophy and is clearly concerned with central philosophical questions, such as the relation of mind to body and the origins of concepts and knowledge.
4. How is artificial intelligence related to psychology? On the one hand, suitably programmed machines seem to share cognitive abilities with organisms and to be useful models of minds. On the other, digital computers are merely syntactic engines, where artificial intelligence may be to intelligence as artificial flowers are to flowers.

In addition to these interdisciplinary problems, the overarching philosophical problem in cognitive science is the venerable one of the relation of the mental to the physical. How can physical systems have mental properties? The *identity theory* develops the suggestion that mental events are just a subset of physical events. However, if these mental events and specific physical events are one and the same, 'they' must have the same properties. But philosophers have often noted that mental and physical events seem to differ: minds are in contact with the eternal, everything physical is changing and ephemeral (Plato); minds are unified and indivisible, physical things are in principle infinitely divisible (Descartes); mind is creative, matter is not; physical events are located in space, mental events are not (e.g., Shaffer).

Defenders of the identity theory may present various replies to these objections: they may assert that, in whatever sense the mind is in contact with the eternal, a changing physical system may enjoy such contact for a time. An 'ephemeral' pocket calculator can *embody* eternal mathematical truths.

Similarly, it may be asserted that creativity is compatible with the nature of the physical. This is fairly clear to the extent creativity involves novel recombinations of preexisting elements. As the example of biological evolution shows, extensive and diverse creation does not require the supernatural.

These defenses involve denying alleged limitations of the physical. With regard to the assertion that the mental is inherently unified, however, the tack is the opposite: materialists can deny that the mind is indivisible. Since Freud, it has been widely accepted that much of the mind is not available to introspection; the mind thus involves distinct parts or realms. Recent cognitive psychology completely supports this general finding; the mind is better conceived as an interconnected system than as a monad [cf., e.g., Fodor (1983)].

All of the preceding replies involve asserting that some property in which the mental and the physical allegedly differ is really present in both the physical and the mental. The situation may be more complicated, however, with regard to the claim that physical events all have location whereas mental events do not. One may assert that mental events actually do have location, that thoughts literally occur in one's head. On the other hand, one may deny that *either* physical *or* mental events have location. If events are analyzed as abstract entities, say, as ordered pairs of space-time slices or as ordered *n*-tuples of objects, properties and times, then no event has location. On such analyses, the objects participating in an event may have locations, but not the event itself. A funny thing may have happened on the way to the Forum, but no *event* was on its way to the Forum. This reply thus asserts that mental events can be physical events because neither has location.

An enduring problem for materialism generally and the identity theory especially is, how can physical events possibly have qualitative character? In particular, for any neuronal events, could it not be the case that those events take place and yet there be no subjective experience that accompanies them? Let us call this possibility 'the missing qualia' objection to materialism. Clearly, if it is possible that physical event *A* exist while mental event *B* does not, then *A* cannot be identical with *B* (that is, they cannot be one and the same event). But for any suggested identity between mental and physical events, it does seem possible that the physical occur but the mental not – the missing qualia objection. Something very like this concern is pushed by Leibniz:

Moreover, it must be avowed that *perception* and what depends upon it *cannot*

possibly be explained by mechanical reasons, that is, by figure and movement. Suppose that there be a machine, the structure of which produces thinking, feeling, and perceiving; imagine this machine enlarged but preserving the same proportions, so that you might enter it as if it were a mill. This being supposed, you might enter its inside; but what would you observe there? Nothing but parts which push and move each other, and never anything that could explain perception [thought]. (Leibniz, *Monadology*, Section 17)

Not only does this problem have an extensive history, it also cuts deeply into the core of cognitive science. At issue is not just the ultimate abilities of digital computers and the role that AI might play in producing an understanding of the mind – let alone, in producing a mind. At issue is the nature of mind. Can the mind be described at some level as a purely computational device, a symbol manipulating engine that operates only on syntactic features of inputs and its own internal representations? Or could it be that in such a device there would, as Leibniz says, “never be anything that could explain” thought and sentience?

An attempt at an end-run around these concerns was the Turing Test. In the spirit of behaviorism, Turing suggested that if a machine was indistinguishable in linguistic behavior from a human, it was intelligent. It is clear that there are several philosophical assumptions implicit in this test. The Turing Test focuses attention on two aspects of the mental to the exclusion of all else. First, concerning behavior, it requires merely that intelligence be simulated. Thus it presupposes that there is no difference between simulated intelligence and intelligence.

Second, the Turing Test takes *linguistic* behavior to be the sole indicator of intelligence. This was probably originally for convenience, as I/O devices are commonly best suited to linguistic communication, and also because the test was intended to provide sufficient but not necessary conditions for intelligence. However, these same factors result in a peculiarly Dr. Doolittlesque emphasis on language in a universe of organisms where intelligence is commonly exercised in arenas other than communication. While humans are much more reliant than myriad other species upon communication, most humans spend little time communicating. One can imagine an alternative to the Turing Test with the same goal, i.e., of operationalizing artificial intelligence, which, instead of a computer terminal connected to a machine or a person, featured a mobile box on castors with a viewing port and air grill at one end. The tester, without approaching closer than, say, five feet, is to determine whether the box is a robot or instead contains a dog. Clearly, this test emphasizes perceptual abilities and appropriate movement. Whether it would be harder or easier to pass is difficult to answer.

An important contemporary philosophical position is that the Turing Test is in principle inadequate. A now classic modern argument that no computer can ever literally understand language or have other psychological qualities that involve meaning is given by John Searle. Unlike Leibniz, Searle is a materialist who believes that physical systems can have mental properties. But he believes that only living biological systems can have minds. Searle argues that there is a real distinction between intelligence or understanding and a simulation of intelligence or understanding. The Turing Test fails to be sensitive to this distinction.

Searle has centered his argument around a scenario now known as The Chinese Room. Searle's argument also counts against the adequacy of the Turing Test: that a machine passes the Turing Test is no reason for supposing that it is intelligent or understands language. Searle asks that we consider a human simulation of a computer running a program which allegedly produces understanding of a subset of English, such as those produced by Roger Schank that employ 'scripts'. Imagine a human being who has not learned Chinese who sits in a room with a vast number of instructions written in English. The complicated instructions tell the occupant of the room to wait until a piece of paper is slipped under the door of the room, and then to look at the marks and shapes painted on the paper. Then, by following many complicated instructions involving intermediate steps and extensive record-keeping in vast ledgers, the occupant paints certain brush strokes on another piece of paper and finally pushes his artwork back under the door.

Unbeknownst to the occupant, the original piece of paper was inscribed with characters of the Chinese language, and the artwork he produced also consists of Chinese characters that are interpreted by persons outside the room as *answers* to the *questions* they had slipped under the door of the room. Searle concludes that, however clever the instructions are and however appropriate the output of the process is as response to questions in Chinese, the occupant of the room does not understand or come to understand Chinese. The occupant does not know what he is doing and does not know the meaning of any Chinese words. The Chinese room might pass a Turing Test, but there is no understanding of Chinese there. Thus it is possible to produce excellent simulations of understanding, but no matter how good they are they will not be understanding.

From the standpoints of the linguist, the logician, and the philosopher of language, the problem of explaining the relation of mind to symbol manipulation is reflected in the problem of the relation of syntax to semantics. On the face of it, syntax and semantics appear to be two quite distinct aspects of

language. Syntax involves the relation of symbols to each other, comprising formation and transformation rules: grammar, broadly construed. Syntax apparently sets bounds on what *can* be meaningful, providing criteria for distinguishing well-formed potential bearers of meaning from ill-formed nonsense. Semantics involves the relation of symbols to the world.

Some logicians and philosophers of language eschew epistemological concerns (or embrace neoplatonistic epistemologies) and construe the world as containing certain timeless abstract entities such as sets and truth-values (extensional semantics) or properties and propositions (intensional semantics), in addition to concrete space-occupying things. There are advantages to such ontological inflation: the primary semantic relation between language and the world is that of reference. And the truth of a sentence or a statement is a function of the reference of its constituents. For example, on an extensional account the truth of a sentence of the form of ‘*a* is *F*’, where ‘*a*’ is replaced by a name and ‘*F*’ by a predicate, is determined by whether or not the item referred to by the name is a member of the set referred to by the predicate. The general result – that the truth conditions for the whole are determined by the meanings of the parts – is highly desirable. It satisfies additional epistemological constraints on semantics: a language user can understand (know the meaning of) an infinity of sentences by knowing the meaning of a finite set of constituents. Indeed, Donald Davidson has suggested that this is a necessary condition for any account of semantics.

On these traditional theories of semantics, semantics is an aspect of language quite distinct from syntax. Syntax provides only formal operations on meaningless symbols. Meaning requires the *interpretation* of the symbols: meaning must be assigned to them. This in turn requires an interpreting *mind*, an *interpreter* who can construe or understand the symbols as meaning one thing rather than another. But then there can be no ‘emergent’ semantics that is generated by syntactic operations of suitable nature and complexity. No system of purely syntactic operations can pull itself up by its own bootstraps to become a mind that understands and ascribes meaning.

These philosophical considerations would appear to have large implications for linguistics, for work in computer science on artificial intelligence, for computational theories within cognitive psychology, and even for neurophysiology. The consideration that semantic properties cannot emerge from syntactic operations cuts two ways. If a system has semantic properties (intentionality) then so must its parts: it must have semantic properties at the lowest level at which a complete description can be given. Contrapositively, if a system does not have semantic properties at the lowest level at which it

can be described, it cannot have them at any higher level of description. In particular, if some form of materialism is correct, such that human minds are in some sense no more than neuronal activity, then given the preceding strictures on the possibility of emergent meaning, the functioning of neurons cannot be purely syntactic. It is just as inappropriate to describe mental operations at some higher remove from the physical (as in cognitive psychology) as purely computational. Indeed, given that, unlike neurophysiology, the basic operations in classical computers are well understood and are purely syntactic/computational, no computer will ever understand language or possess genuinely mental characteristics. And linguists must see that generative grammar *per se* cannot introduce semantic properties except as a distinct interpretation of syntactically acceptable entities.

There is no generally accepted solution to these difficulties, a fact reflected in the central place occupied in this collection by often opposed computational, connectionist, representational, intentional, and dispositional conceptions of mind. It is clear that despite the problem of the gap between syntax and semantics, computational conceptions of mind have not been abandoned, nor have workers in artificial intelligence abandoned pretensions to be mind creators. A sketch of why this should be the case might therefore be in order.

The formal systems of the logician are totally unconnected from the world and hence surely do require an interpretation if they are to represent anything. The question arises whether the intervention of a mind is the *only* way in which a syntactic system can come to have semantic properties, i.e. can come to represent, as is presupposed by the philosophical argument presented above. A suggestion that there might be alternative sources of meaning comes from various theories which have revolutionized philosophy over the last twenty years. What these theories in diverse domains all share is a central position accorded *causality*. In philosophy of language there is what is variously called 'the new theory of reference' or 'the causal theory of reference'. On this account, terms in ordinary language come to refer as they do in virtue of causal connections between a current user of a referring expression and its referent. These causal connections determine the meaning (the exact reference) of the referring expression. The meaning is not determined by any idea in the mind of the user. This account thus diminishes the importance of the user as an interpreter in bestowing meaning on the terms he or she uses. (The alternative which made the user all-important was sometimes disparagingly called the 'Humpty-Dumpty' theory of meaning, after that egg's proclamation that the words he used meant whatever he wanted them to mean.) The causal theory of reference is hardly a complete account

of the causal connections required to give reference, but it is a major new philosophical theory that provides an alternative view of at least one semantic property, and which downplays interpretation in favor of largely external causal connections.

Causality has also been seen to be important in the analysis of perception: what makes *x*'s seeing *y* a seeing of *y* is the causal role played by *y* in this event, rather than what *x* takes *y* to be. Again, this results in diminution of the role of the interpreter. In epistemology, causality appeared to be important in analyses of knowledge. Consideration of cases presented by Edmund Gettier – where someone has a justified true belief yet intuitively fails to have knowledge – suggests that one deficiency in these cases is the absence of an appropriate causal connection between the facts and the person's beliefs about the facts: the fact making the belief true is not the one that caused the belief.

Causality also plays a central role in the now widely accepted *functionalist* theory of mind. This theory was first introduced as a version of materialism. However it avoids problems of reduction by characterizing mental states and events in terms of their connections with stimuli, behavior and other mental states and events. A seminal statement of the theory is D. M. Armstrong's *A Materialist Theory of the Mind*:

The concept of a mental state is primarily the concept of *a state of the person apt for bringing about a certain sort of behaviour*. Sacrificing all accuracy for brevity we can say that, although mind is not behaviour, it is the *cause* of behaviour. In the case of some mental states only they are also *states of the person apt for being brought about by a certain sort of stimulus*. But this latter formula is a secondary one. [Armstrong (1968), p.82]

Armstrong goes on to note that a complete account of mental states should involve not only their causal relation to behavior, but their causal relation to *other mental states*. The account of some mental states must be exclusively in terms of the other mental states that they cause. As an example, Armstrong cites the intention to add 'in one's head': "The intention is a mental cause apt for bringing about thoughts that are the successive steps in the calculation" (*op.cit.*, p. 83). The role played by internal states in Armstrong's functionalism is a clear break from behaviorism.

Armstrong also differs from the behaviorist in his conception of disposition. The behaviorist, such as Gilbert Ryle in his very influential 1949 book *The Concept of Mind*, gives a phenomenalist or reductionistic account of dispositions to behave one way or the other. In contrast Armstrong gives what he calls a 'Realist' account of dispositions.

According to the Realist view, to speak of an object's having a dispositional property entails that the object is in some non-dispositional state or that it has some property ... which is responsible for the object manifesting certain behaviour in certain circumstances ... It is true that we may not know anything of the nature of the non-dispositional state. [Armstrong (1968), p. 86]

Armstrong uses the example of glass being brittle. Brittleness involves structural features (we may not know exactly what they are) that make glass disposed to break, rather than deform, when stressed. As important as the development of a functionalist account has been for modern materialism, not all have been especially pleased to abandon behaviorism and eliminative materialism.

PROLOGUE: PROBLEMS OF MIND

A good brief history of recent philosophy of mind and a survey of current issues is provided in our first selection by one of the leading philosophers in this area, Daniel Dennett. Starting with a description of Ryle's *The Concept of Mind*, Dennett traces the rise and fall of ordinary language analysis, a fall which parallels a corresponding decline in the domination of psychology by behaviorism. Dennett views Hilary Putnam and Jerry Fodor, who are both contributors to this collection, as two of the principal authors of the movement away from both ordinary language analysis and behaviorism. Dennett then describes the advocacy of materialism in the form of mind/brain identity theory in the later 1950s by a group of Australian philosophers who apparently never felt the full civilizing influence of Oxford analytic philosophy. Dennett explains how and why this materialism gave way to the now dominant position, functionalism. Dennett sees functionalism as arising from materialism, but he also notes that it is "a spiritual descendent of logical behaviorism" – in fact it is a neat synthesis of behaviorism and materialism. This surely accounts for the breadth of the appeal of functionalism in current philosophy of mind. The concept of functional role acknowledges both the psychological importance of input-output relations *and* the need for internal, ultimately physical, mechanisms. The papers by William Bechtel and Patricia Kitcher in this collection discuss issues relating to functionalism and the relation of mind to brain.

Dennett concludes with a discussion of current problems. A subsidiary problem concerns epistemology and psychology, i.e. 'incorrigibility' or privileged access – the extent to which each subject is in a special position with regard to knowledge of his or her own mental states. This issue was of

more concern when behaviorism enjoyed more prominence than it does now. The most recent discussion of problems relating to epistemology and cognition concerns the extent to which epistemology can be seen as a subpart of psychology – that is, the extent to which epistemology can be ‘naturalized’, to use W.V.O. Quine’s phrase. That this is the current focus of discussion in this area is reflected by the papers in Section V.

A principal current problem discussed by Dennett is the question of how adequately functionalism can account for the inner conscious experience of a sensation, such as pain or the subjective difference between red and green. Kitcher also discusses this question in her paper, as does David Cole (1990). The other main current problem Dennett discusses is the relation of internal representations to meaning. Dennett cites Roderick Chisholm’s early contribution to this problem. A more recent paper by Chisholm on the locus of meaning may be found in Section IV. The problem of meaning is closely related to the question of the possibility of machine mentality. Accordingly it is a central topic of current discussion. Dennett’s introductory discussion of these problems sets the stage for the papers by Fodor, Searle, Fred Dretske, Smolensky, Ramsey, Stich and Garan, Kenneth Sayre, and James Fetzer that follow.

I. COMPUTATIONAL CONCEPTIONS

Computational theories of mind hold that minds are computational systems (implications for cognitive psychology) and, as a corollary, that suitably constructed computational systems will be minds (implications for artificial intelligence). Computational processes are those that occur solely in virtue of syntactic features of states or inscriptions, and not in virtue of semantic properties (meaning, having to do with what the symbol stands for) or pragmatic properties (how a symbol *user* intends or interprets the symbol). Computational processes have the advantage that their realization by a physical process is relatively straightforward; there is a minimal residue of interpretation, and it is a residue which in many cases we (engineers, that is) know how to accommodate. The syntactic properties of a symbol are relatively close to, but are not, physical properties. That this quotationally displayed mark ‘a’ is an instance of the letter *a* involves interpretation, involving assimilation of the mark to a class of a certain kind. Non-contextual differences between instances of the same syntactic type may be syntactically (and semantically) irrelevant – an ‘a’ is an ‘A’ is an ‘a’. Only those who construct ransom notes from clipped letters take full advantage of this latitude.

Let me sketch a characterization of the analog/digital distinction that is close, on the one hand, to the original unphilosophical engineering distinction between continuous and discrete state machines, and, on the other hand, to the syntax/semantics distinction. For natural language, syntax is defined relative to semantics, contrary to the procedure of logicians (and perhaps certain linguists they inspire). If a physical difference between symbols makes a semantic difference, then *ipso facto* there is a syntactic difference between the symbols as well. But it is not the case that every semantic difference is reflected (surface) syntactically – otherwise there would be a distinction without a difference. From the standpoint of the engineer, every machine is, unfortunately, continuous. But in a digital machine, this continuity is an irksome fact to be minimized and ignored. That is to say, all voltages from 3 to 5.5, say, are syntactically identical – they are treated the same, have the same causal powers, in the system. Voltages from 1 to 2, while they occur of course, are not permitted to be causally efficacious – they are not constituents of well-formed signals. In an analog system, by comparison, every variation in the symbolic magnitude is significant.

On this account, whether a system is analog or digital depends upon the syntax of the system, and that must reflect the semantics. Thus, oddly, the digital/analog distinction is ultimately relative to purposes, what it is taken to represent. Consider an ordinary mechanical ‘grandfather’ clock, weight driven with a pendulum escapement. Such clocks tick, and the instantaneous angular velocity of the hands varies. Such a system can be said to be analog or digital only relative to the meaning of the position of the hands. If the position of the hands is taken to represent the count of pendulum oscillations, then it is a digital system. Slight motion of the hands which occur during a pendulum swing have no semantic significance and do not count as an increment/decrement of the count. If the position of the hands is taken to represent the position of the weight, however, the system is analog. Even very small movements of the hand may represent movement of the weight and will, if the machine is operating perfectly, with inelastic weight support, etc. On this account, then, a sticky clock with hands that shudder from position to position may be analog, although its hands only occupy discrete positions. It is an analog device with limited resolution. Whereas a cash register that has a ‘cents’ dial which very slowly rolls from ‘0’ to ‘1’ as a penny is rung up is nevertheless digital.

To represent, a system must represent something *as* somehow. Why isn’t a thermostat a representational system? I think it is because there is no principled way of distinguishing its states as representing the temperature of

the room, or that heat flow out has exceeded heat introduced, or that its bimetal strip has distorted, or that the furnace is on, or that the world energy supply is being reduced, and so forth. When the thermostat is operating properly, all these are the case whenever it enters its 'on' state. The problem is that the system can make so few discriminations. But this suggests that the difference between representational and other systems may be a continuum.³

Many philosophers, however, take the difference to be deeper and more devastating for the prospects of artificial minds. It is thus appropriate that the first of the selections in this section is Fred Dretske's 'Machines and the Mental'. Dretske is sensitive to Searle's concerns about the adequacy of the Turing Test and the inadequacy of purely syntactic manipulation of symbols as a basis for mentality. But he does not believe that computers are eternally doomed to mindlessness.

Dretske attempts to determine which of a computer's disabilities are such that it cannot think, understand, and so forth:

[Machines] don't do what we do – at least none of the things that, when we do them, exhibit intelligence. And it's not just that they don't do them the way we do them or as well as we do them. They don't do them at all. They don't solve problems, play games, prove theorems, recognize patterns, let alone think, see and remember. They don't even add and subtract.

Computers are artifacts that *we* use when *we* do things: the *artifacts* don't *do* these things, *we* do. Dretske suggests that keys do not open locks; people open locks by using keys. That seems right – because keys cannot open locks by themselves. You have to hold them, insert them, and turn them. The disability of the key is not that someone uses it; it is that it can't get the job done without help. A key is passive. Automatic devices, on the other hand, can do things by themselves, and computers are automatic devices. But this is not enough, Dretske argues, to tell us just *what* a computer is doing automatically:

To understand *what* a system is doing when it manipulates symbols, it is necessary to know, not just what these symbols mean, what interpretation they have been, or can be, assigned, but what they mean to the system performing the operations. ... One thing seems perfectly clear, if the meaning of the symbols on which a machine performs its operations is a meaning wholly derived from us, its users – if it is a meaning that *we* assign the various states of the machine, and, therefore, a meaning that *we* can change at will without altering the *way* these symbols are processed by the machine itself – then there is no way the machine can acquire understanding, no way these symbols can have a meaning to *the machine itself*.

What is therefore required is to provide a direct connection of the computer to the world – as a robot with sensing devices, perhaps. Dretske contrasts the case of the computer, typically fed symbols by humans via keyboard, with that of a living creature, which internally represents meaningful – indeed life or death – information about its environment. Dretske thus appears to disagree with Searle on ‘The Robot Reply’, since Dretske suggests that the lack of meaning to the machine of the symbols it manipulates might be remedied by connecting the computer to real-world sensors. Dretske concludes that work on machine perception, pattern recognition, and robotics has greater relevance to the cognitive capabilities of machines than even sophisticated programming of such purely intellectual tasks as language translation, theorem proving, or game playing.

Dretske considers the limitations of lower animals that motivate our tendency to decline to attribute to them specific beliefs. A police dog may be a good *detector* of marijuana, Dretske notes, but the dog does not respond to the smell in the way it does because of what the smell *means* (that there is an illicit drug – marijuana – present). Rather, the dog has merely been conditioned to respond differentially to a smell without knowing what it *means*. The dog responds in the way it does, not because it knows it has detected marijuana, but because it previously has been rewarded for responding to that smell. Similarly, we would not attribute much in the way of beliefs to artificial detectors (to smoke detectors, for example, or to computerized detectors) if the meaning of what they detect does not play an essential role in the production of the response.

In many respects Dretske’s argument is similar to Searle’s. But there is an important divergence in the character of their conclusions. Searle presents an *in principle* objection; if he is correct computers are *intrinsically* or essentially incapable of understanding natural language. Dretske presents *in practice* or contingent objections to the state-of-the-art and certain ways of going about the implementation of artificial intelligence. He holds that no ordinary immobile computer, without sensory input and without effectors other than CRT screens and printers, can understand. Dretske’s objections thus might be overcome by adding sensors and effectors.

While Dretske’s position is certainly thought-provoking, it is not clear how to appraise the force of his objections with reference to current practice. Despite what he says, it is not clear why a machine needs to be directly connected with its environment in order to add or understand. Since in our time computers do almost all routine calculations, it is easy to overlook that not so long ago human clerks were employed by business and by government

to perform routine addition. Clerks need not know what the symbols represent in order to add them – or indeed whether they represent anything at all. So it is not clear that Dretske is correct in claiming that computers do not add merely because humans interpret their input and output.

Dretske introduces two conditions for performing certain symbolic operations (mental acts?). The first of these is that “the symbols must mean something to the system performing the operations” (p. 80). But why should we suppose this is true of anything other than public language? It would be a mistake, I think, to conflate the language used for communication with the representational system used for thought. I, after all, would not know what one of my brain states meant if you described it physically or showed me a 3-D micrograph of it. I do not construe or interpret my brain states; they do not appear to me as symbols. Yet my brain states represent the world *for* me, but not *to* me. As Dretske seems to concede, “Brains have their own coding systems … In this respect a person is no different than a computer.” (p. 78–9)

Dretske’s second condition is that the computer must have *information* in order to use symbols that have meaning. As I argue above, however, the symbols manipulated internally need not mean something *to* the machine. Dretske holds that for a machine to have mental states, the meaning of symbols processed by the machine must not depend upon the interpretation of the human *user* of the machine. I believe that this is not quite correct, although Dretske is on the right track. For Dretske seems to think, correctly I think, that a robot could use meaningful symbols because of connections the symbols have with the world. Here Dretske rightly parts company with Searle. The question is whether the presence of a human being in the causal chain which connects symbol with world is detrimental to the information born by the symbol.

Perhaps a hypothetical scenario can best illustrate the objection. I can see and I glean information about the world by using my eyes. Suppose it were discovered that living on the back of my retina was a tiny homunculus who, in me, occupied the place normally held by a neuron. The missing neuron is one that normally transmits information from a rod cell in the retina to the optic nerve. When the homunculus determines that the rod cell is stimulated, it interprets this correctly and squirts a tiny bit of neurotransmitter into synapses exactly as the missing neuron would have done. That is, suppose that this homunculus is functionally equivalent to a neuron. Now it seems to me that the presence of the homunculus would have no effect on my sightedness. The visual information I glean from the world would not be degraded by this intervening mind. Thus it is not clear that, when interpreting humans

act as eyes and ears for a computer, the computer is prevented from understanding as a result. Humans can form a link in the causal chain which connects the computer with the world. As a result, Dretske's argument places a great deal of weight on the distinction between *direct* and *indirect* connections with the world, but it is far from clear that this distinction can bear the weight of marking a difference between having and not having mental states.

In 'What's in a Mind?', Zenon Pylyshyn notes that, however philosophically appealing the project of constructing causal theories in psychology may be, the project has turned out to be difficult. He argues that this is not due to the complexity of organisms, but to uncertainties about the domain of psychology itself, about what psychological theories should do. Pylyshyn argues that mistaken preconceptions about the basic structure of mental phenomena continue to impede successful explanation. For example, treating the domain of psychology to be just that of conscious mental contents has rather notoriously failed to produce good psychology. Pylyshyn argues that it is not primarily for methodological reasons that this is so (e.g., that conscious contents are not public or are perniciously altered by the act of observation itself), but rather because conscious contents are not a natural kind, not a self-contained domain of explanation containing both explanans and explananda. On this basis, Pylyshyn criticizes phenomenological critiques of AI advanced by Rudolph Arnheim, Hubert Dreyfus, and S. M. Kosslyn. In doing so, he also places himself squarely in opposition to the defense by John Searle of intentionality as a natural phenomenon. Searle's views are discussed below.

Pylyshyn's main interest in this piece is in distinguishing two kinds or levels of explanation of the behavior of a system: in terms of the intrinsic mechanisms or functional architecture of the system, and in terms of manipulation of *representations*, where the content of the representations (semantics) plays a crucial role in the explanation. Pylyshyn says that the latter behavior is nomologically arbitrary because these regularities are not subsumed under natural laws, and he cites rules of grammar and inference as examples. Pylyshyn goes on to argue that some regularities in linguistic behavior are best explained by appeal to tacit knowledge of orthographic rules rather than by appealing to the open-ended intrinsic capabilities of the system. The system itself is capable of an enormously wide range of behavior (e.g., humans in principle can speak any of an infinity of possible languages). The behavior of the system is constrained by information (beliefs, knowledge), where these must figure in explanations of its behavior. Pylyshyn argues more fully for this position in *Computation and Cognition* (1984).

Pylyshyn claims that any difficulties in developing a successful explanatory paradigm in psychology are not because psychology is young – he notes that it is as old as physics. But this may be a bit too facile and violates, at a metalevel, Pylyshyn's own concerns about what we view as natural kinds in psychology: it treats speculation and research programs having to do with 'the mind' as all of a type. If we take a finer grained look at the enterprise of psychology itself (not its domain, as Pylyshyn is doing), there are clearly many quite different approaches that might retrospectively be called psychology, most of which are not as old as physics. The current causal approach in psychology is indeed young; like any young science, it is long on prospect and short on demonstrable successes. But a comparison with the interval between the proposal that there are biochemical explanations of life processes, originally offered by the Presocratic atomists, and the first clearly successful specific biochemical explanations of those processes, should temper our expectations of what, say, passage of a mere two thousand years might bring about.

II. CONNECTIONIST CONCEPTIONS

Connectionism is a research project and paradigm in cognitive science which has recently received a great deal of attention. There are several aspects of the program, and at least three positions which might be connectionist. Viewed as an area of theoretical investigation, parallel distributed processing is an arena of intrinsic interest as a type of possible information processing system unlike those with a classic centralized von Neumann architecture. Research can explore the capabilities of such systems and attempts can be made to devise efficient and productive designs as well as search for theorems regarding the nature of these systems. There need be no claim that this is a description of any mind which happens to exist as an historical accident of the history of evolution. Nevertheless, this investigation is both mathematical and empirical. While in principle, perhaps, one could describe the properties of parallel distributed processing systems completely *a priori*, in practice it is necessary to implement systems and see what they do. Let us call this position theoretical (or 'hyper-weak') connectionism.

However, connectionism, as the '-ism' might suggest, also has a metaphysical aspect. The program is typically based upon claims about the actual character of minds, animal and human. Interest in parallel distributed processing developed not simply as a theoretical model of a *possible* way of constructing information processing systems, but one with a wide-open eye

on the actual character or biological systems at the level of neurophysiology. The units of connectionist systems behave, with regard to their interaction, much as neurons or small clusters of neurons behave. They can be viewed as standing to neurophysiology as the frictionless plane is to mechanics. It seems clear that at some level minds are realized by something like the systems of connectionism. Neurons just are interconnected loci of activity which are triggered into activity and which contribute to the activity of other units.

Let us call the position that actual minds are describable by connectionist models at an abstract functional level once removed from the physical as 'weak' connectionism. This view is consistent with the position that actual, biological, minds are also describable at some higher level as implementations of rules and representations. Indeed, this position appears to be held by McClelland, Rummelhart and Hinton in their manifesto, 'The Appeal of Parallel Distributed Processing'. They appear to hold a compatibilist position, maintaining that actual minds are fundamentally connectionist but that they can still have rules and representation descriptions. But some debate exists (see below) as to whether or not the connectionist system *implements* the rules and representations system.

Connectionism is not motivated solely by a bottom-up look at the physical underpinnings of mentality. Investigations of pattern recognition and associative memory at a higher functional level of information processing converge to suggest that nodes activate other nodes and that there is not generally a centralized process with well defined abstract syntactic rules that describe what actual organisms do. Thus, the rules of generative grammar may characterize human potential, but they do not approximate how we actually produce what we do achieve. And as we move away from the abstracted and processed to the sensory – from, say, the recognition of grammaticality of well-defined formulae to real auditory speech recognition – the artificiality of classical artificial intelligence becomes more apparent. And if the proof is in the pudding, practical general speech recognition computer capabilities remain unavailable, despite the incentives of potential billion dollar markets. Let us call the view that minds can *only* be realized on a parallel distributed processing, as an exclusive alternative to rules and representations, 'strong' connectionism.

Given the known physiological facts, it is difficult to doubt that at some level a connectionist approach may be correct. Thus, at a minimum, weak connectionism appears to be true. The interesting question at hand concerns where the correct level of connectionist description is to be found. In part the

debate is between those who believe that connectionism shows that the classic rules and representation approach to mentality is completely wrong, and those who think that it is correct. The latter believe that connectionism turns out to be fully compatible with a rules-and-representation approach. Such a position acknowledges that weak connectionism may be true but contends that other important projects must be pursued in cognitive science. Furthermore, it is held, for example by Fodor and Pylyshyn, that the rules-and-representation approach remains essential to do justice to the inferential capabilities of human minds.

In the selections on connectionism that follow, the first view – that connectionism represents a radical departure from classic models – is clearly and forcefully set forth by William Ramsey, Stephen Stich and Joseph Garan. They are careful to qualify their view with the big ‘if’: if connectionism is a correct description of actual organisms, then it shows that classical approaches are incorrect – and that the advent of connectionism is a conceptual revolution on a par with the Copernican revolution.

A more moderate view is presented in the influential paper by Paul Smolensky, ‘On the Proper Treatment of Connectionism’. Smolensky presents a useful overview of the connectionist approach, along with an interpretation of its status as a theory of mind. Smolensky’s moderate interpretation may be becoming a consensus view of workers in the area. The central philosophical question has to do with the status of connectionism with regard to two traditional levels of description of organisms with mental capacity: a high level description involving conscious symbol manipulation, and a low level description of the physical operation of the brain. Smolensky argues that connectionism introduces a new paradigm of description, the *subsymbolic*, which, although perhaps close to the symbolic, is not identical with it and in fact shows it to be at best an approximation of actual mental activity. Thus connectionism is a genuine alternative theory to both neurophysiological as well as rules-and-representation accounts of the mind.

Smolensky holds that the subsymbolic processes characterize all mental operations. Smolensky draws on two distinctions: one between conscious and subconscious mental operations, and another between individual and social knowledge. His position is that conscious thought processes do involve rules and conceptual symbols. But these, in an important sense, do not reflect the basic character of mind. The rules and concepts used in conscious thought are an overlay on a subsymbolic system. Conscious mental activity is activity of a virtual machine approximately realized by the underlying subsymbolic system. Rules and concepts permit social knowledge to be formalized and

communicated and then to be used, in the absence of first-hand individual experiential acquisition. Nevertheless, they are artifacts and should not be seen as constituting the essence or bulk of cognition.

Most of Smolensky's paper is concerned with carefully laying out the character of the subsymbolic connectionist system. The theoretical problem is considerable: the subsymbolic system must be distinguished from both the neural and the conceptual. According to Smolensky, the subsymbolic system *is* representational. That is, states of the system have the semantic property of representing features of the environment. What then distinguishes the subsymbolic from the conceptual? Unfortunately, the answer is not completely clear. In particular, it is not clear why the dynamic entities in the subsymbolic system should not be regarded as alternate concepts. It may be that the elements of the subsymbolic system that represent familiar features of the environment, such as other people, animals, and the physical objects that are characteristic of the everyday world are not identical in semantic properties with those of a publicly communicable concept system. But this does not show that they are not symbols or concepts. Could one regard them simply as an alternative, private, fluid protoconcept system, bearing a family resemblance to Fodor's language of thought? At some point, after all, on Smolensky's account the subsymbolic system must realize the conceptual rule-following system. If we are to take the talk of virtual machines seriously, then it would seem that the system must be capable of producing all the operations of the virtual machine. Thus one way of viewing connectionism is as an implementation of rules and representation systems.

What seems especially interesting is that connectionism might afford an alternative to Fodor's strong nativism hypothesis. Fodor seems to suppose that there is an inborn primitive set of concepts sufficient to represent all concepts ever acquired. Connectionism, I think, may offer an account of how a highly plastic system can acquire concepts in virtue of training. It might solve the bootstrapping problem that motivated Fodor's nativist position.

Thus defenders of the traditional rules and representation approach may view connectionism as merely an implementation. But Smolensky rejects this view. His remarks, amplified in his 'The Constituent Structure of Connectionist Mental States: A Reply to Fodor and Pylyshyn', suggest important problems for any straightforward interpretation of the relation between the rules and representation system and the connectionist system as implementation. Smolensky draws on a suggestive metaphor, that of the relation of classical mechanics and quantum mechanics. Classical mechanics is an approximation, and, strictly speaking, is false. As I interpret him, Smolensky

is making the same claim about the rules and representation account. Unlike a von Neumann computer, the brain only approximately comes to implement rules and representations. The basic structure of neural systems is not characterized by primitive rules which can be compositionally structured so as to perfectly realize high level rules. The connectionist approach shows how a 'soft' system of sufficient complexity can appear 'hard', that is, approximate a classical rules and representation cognitive system. If this is an implementation, it does not appear to be a classical implementation.

Against both the positions of Ramsey *et al.* and Smolensky the classic perspective has been prominently defended by Fodor and Pylyshyn. They view connectionism as a view from the bottom up, contending that capturing the higher logical capabilities of minds requires the classical representations and rules approach. They are compatibilists, who allow the possible truth of weak connectionism. But they deny the zeal of would-be strong connectionists; for even if connectionism is a correct abstraction from the neurophysiological (so that weak connectionism is plausible), the rules and representation approach remains the correct conception of actual minds.

Ramsey, Stich and Garan argue that connectionism if true would establish that folk psychology *misd*escribes the mind: if connectionism is correct, it is strong connectionism that is correct. There is nothing in connectionist systems corresponding to beliefs and desires. For even in moderately complex connectionist systems – to the extent that it makes sense to speak of the *representation* or embodiment of information at all – the information is embodied diffusely. Thus Stich, in particular, a long time critic of the adequacy of ordinary or folk psychology, takes comfort in the prospect that connectionism may become the accepted paradigm. For with its acceptance, belief and desire will be driven from the field of acceptable constructs in cognitive psychology.

It may be that the authors of the two papers have a fundamental difference about the character of representation. Ramsey *et al.* hold that nothing in connectionist systems represent and Smolensky holds that connectionism is a representational view, and that it does not afford comfort to eliminativists. Perhaps the difference turns on *what* represents. It seems clear that not every node need express some property or state that one can conveniently label. But critics such as Fodor and Pylyshyn apparently hold that some must, or that aggregates of nodes or states of the entire system must represent. However Fodor and Pylyshyn note that what I have called strong connectionists deny that the representations have a 'combinatorial constituent structure' which would permit the application of rules sensitive to such a structure. Fodor and

Pylyshyn contend that many connectionists have erroneously supposed that on the classical view the rules themselves must have an explicit syntactic representation. As they note, I think correctly, only the data upon which the rules operate need have structured representations. It is an artifact of the need to implement classical modes as virtual machines which requires representation of the rules themselves.

The argument offered by Fodor and Pylyshyn in favor of classical models and against strong connectionism turns on the “productivity, systematicity, compositionality and inferential coherence” of the mind. As it turns out, these considerations are all closely related. Capacities are productive if they can, apart from resource constraints, produce a potential infinity of results. Systematicity appears to be a weaker feature more clearly demonstrable: humans come to have general, if not infinite, capacities on the basis of very limited exposure to samples. Compositionality is required to account for our general understanding of language. And (deductive) logical inference is another general capacity that requires rules operating on structured representations.

It is worth remarking that the evidence marshalled in favor of classical views turns on linguistic competence. As near as I can tell, Fodor and Pylyshyn's arguments leave open the possibility that strong connectionism is true of every animal other than humans. And they are compatible with a strong connectionist account of human capacities that are not linguistic. Even more narrowly, a rules and representation account may be required only for sentence recognition and generation and for deductive inference based on such representation – other processes involved in linguistic behavior, such as associative memory, affect, and speech recognition may be connectionist. Thus just how much turns on whether or not Fodor and Pylyshyn are correct ultimately depends on how central grammatically and deductive logic are to human cognition.

It is possible to discern another possible view, the ‘hyper-strong’ connectionist view, that connectionism is not merely a correct description of actual minds at *all* levels but also that no possible mind *could operate* on the rule and representation model. This position adopts the most negative attitude toward the rules and representation approach, potentially relegating it to the heap of inconsistent theories. This position seems the closest to that of the most severe critics of classic artificial intelligence, such as Searle and Dreyfus. In this collection, related criticisms are advanced in Dretske's paper, ‘Machines and the Mental’, and Searle's positive views about the character of intentionality are presented in his ‘Intentionality and its Place in Nature’.

III. REPRESENTATIONAL CONCEPTIONS

Despite the apparently provincial nature of Fodor's focus suggested by the title, 'Semantics Wisconsin Style' is actually a defense of a naturalistic account of representation. Fodor's goal is to give necessary and sufficient conditions for '*R* represents *S*'. The Wisconsin suggestion is that representation depends upon a causal connection between the representation and the represented.

Fodor dismisses two naturalistic alternatives to a causal theory: resemblance, and an 'epistemic' account. The latter is to the effect that *R* represents *S* just in case a subject can learn that *S* from *R*. Both accounts run afoul of a singularity aspect of representation, according to Fodor. Put briefly, in most cases a representation both resembles and can be a source of knowledge about much more than just that particular which it represents.

A causal account has the clear advantage of being both naturalistic and accommodating this singularity condition. The main problem for causal accounts is that they apparently cannot deal properly with *misrepresentations*, i.e., with falsity. If *R* represents *S* just in case *S* causes *R*, then all representations are correct. To know the content of a belief, say, would be to know that what it purports is in fact the case – to understand would be to know – to a pernicious extent.

Since most anything can cause most anything, given abnormal circumstances or Rube Goldfarb contrivances, the actual causes of a representation cannot generally determine its content. Further, such an appeal to actual causes of belief will conflate meaning and evidence. The obvious strategy is to confine the analysis to a subset of causes, to attempt to get just the right ones. Fodor discusses Dretske's attempt, which appeals to a learning period to fix content. Fodor argues that in addition to failing to afford a basis for content for *unlearned* representation systems, the distinction between 'learning period' and 'period of use' (where misrepresentation is possible) is an artificial one. Finally, he argues that this account will not work. Misrepresentation is impossible because anything that could have caused the representation – including what we intuitively would deny is represented by it – becomes part of what is disjunctively represented, on Dretske's account.

So, with Calvin Coolidge, Fodor advocates a return to normalcy as a more promising way of constraining the range of causes of a representation that count as providing its content or truth conditions. But this strategy will not work in any simple way, Fodor argues, because there is no principled way of distinguishing normalcy conditions from content, apart from an appeal to

intentions. The problem is still the same: on a causal account a representation represents its causes. The appeal to normalcy thus reveals a covert use of intentions and purposes. A mercury column barometer reflects pressure, but also responds to temperature and variations in gravitation. To take it as representing barometric pressure is to appeal to what it is used for, to purposes. Presumably there is some parallel account of normalcy for *innate representations*; here the appeal will be to teleology, i.e. to biological function. It is notorious that a statistical characterization of biological function has counter-examples: the function or purpose of human skin is not to provide a fundus for pimples, despite the frequency of acne; economists may be right less than half the time, but that does not affect meaning of their representations, in the 'nonnatural' sense of meaning at issue here.

Fodor concludes that it is best to embrace an appeal to intentions and go on to provide – or at least expect – a naturalistic account of intentions. Although he does not explicitly say so, it appears that Fodor's account comes down to something like: *R* represents *S* just in case *R* would have been the output of a properly functioning system the purpose of which is to output *R* just in case *S*. For a system meeting all these conditions, of course, misrepresentation is impossible.

In 'Cognitive Science and the Problem of Semantic Content', Ken Sayre presents a criticism of contemporary cognitive science and in particular of Fodor's attempt to characterize the features of a brain state or process in virtue of which it has semantic properties such as reference, meaning or truth value.

Sayre suggests that Fodor's account is based on a confusion between two senses of 'information'. Sayre goes on to develop a positive account in terms of the univocal technical sense of information as it is used in communication theory. At the outset of his discussion, Sayre gives an overview of Fodor's model of cognitive science, a model shared in most respects by many cognitive scientists. Sayre says that this model requires that it be made clear how "...physical computers, of either the mechanical or the biological variety, can deal with meaning-laden formulae in a manner respecting their propositional content" (p. 231). But Sayre and Fodor observe that many of the semantic and referential properties of its representations involve relations to items external to the system itself. The internal operations of the system itself have access only to the syntactic or formal properties of states. (Sayre calls this the 'formality condition' on cognitive science accounts.) As an example, the semantic fact that a particular belief of mine is about Lake Superior is a relation between my belief state and the lake. In the cognitive

science model of Fodor and others, however, my mental processes have no access to the lake itself, and hence not to the semantic properties of my representations. Fodor's solution to this predicament is to argue that these semantic properties can be mirrored by syntactic/causal properties of the representations. What makes the operation of an adding machine an instance of adding, for example, is that the machine is constructed in such a way that its operations conform to the rules of arithmetic. That machines can literally *embody* semantic rules in this way shows that the brain might do so as well.

But Sayre argues that this view is incoherent. In the first place, the internal states of a computer, such as in the running of a theorem-proving program, do not literally have any semantic properties. They could have such properties only under an interpretation, yet there are many interpretations which they might be given. As Sayre points out, one might think of a theorem-proving program as instantiating *truth* preserving rules; but one could just as well construe the states as having other properties, such as warranted assertability. Since semantic properties of the states, if any, depend upon interpretation by a user who is external to the system, this cannot serve as a model of how *brain* states embody semantics as syntax, on pain of an infinite regress. At this point, therefore, Sayre's argument is similar to that of critics of AI such as Dretske and Searle.

This failing of Fodor's account, Sayre goes on to argue, is due to the conflation of two senses of the much abused term 'information'. In one sense – apparently used by Fodor – to bear information is to have semantic properties. Something embodies information in this sense if it represents some state of affairs. On the other hand, information, in the sense used in information theory (Sayre calls this 'info (t)', is just reduced uncertainty, and has nothing to do with interpretation, meaning, or reference. But it is only in this second sense that computers are information processing systems. Thus "the computer model is powerless as an aid to genuine understanding of what semantic content amounts to and how it originates". (p. 241)

Sayre then presents what he calls an outline of a *communication-theoretic* account of representation. An organism's brain state C which represents environmental state E has the functional property that it enables the organism to discriminate E from other possible environmental states that are of interest to the organism, the relevant alternatives to E . Then, paraphrasing Sayre's information-theoretic discussion greatly, brain state C represents E if it reliably bears the information that E , such that the organism can efficiently react to E . As Sayre sums up:

Brain state C functions as a *representation* of environmental state Ec when together with Ec it constitutes a communication channel governed by these two constraints – the interacting requirements of faithful $info(t)$ -transmission from environment to nervous system, and of efficient $information(t)$ -processing on the part of the latter.

Sayre concludes by noting that his account requires that biology play a large role in cognitive science, since the representational properties of brain states depend upon how the brain actually works in processing $info(t)$.

It appears to me that Sayre's exegesis of Fodor is correct. It also appears to me that his information theoretic account of representation is basically correct. What remains, as we so often find in these assessments, is to determine if the condemned party – in this case Fodor – is actually guilty as charged. As Sayre admits, his account does not pay much attention to the *causal* connection between representation and represented, saying “the particular physical connection between C and Ec is not an essential component of that relationship [the relation of representation]”. In the accompanying note (note 32), Sayre says that, “For the most part, however, meaningful symbols are not caused by the things they represent. Similarly, it is not the *causal* aspects of the process by which brain state C is produced that make it a representation of environmental states Ec .”

I find this baffling. In the first place, it is precisely the causal relations between output and input that make a real communication channel instantiate the information-theoretic properties that it has. Secondly – and less technically – this claim reflects a very pessimistic view concerning the verisimilitude of most human representation. I wish to believe that the ‘meaningful symbols’ that are perused in surveying the stock market report or the front page news or the football scores are in fact caused by the things they represent. I recognize that – to a decreasing and perhaps vanishing extent, especially in the case of stock market reports – those causal connections *include* human beings. Interpreting human beings are *part* of the causal chain that constitutes the ‘communication channel’ that connects our brain states with the states of the world. Similarly, interpreting human beings are often but not always part of the causal chain that connects computers with states of the world. So, contrary to what Sayre says, it is possible for computer states to represent states in the world in the absence of an interpreter of those states, because those computer states *are* causally linked with the states of the world.

Like others who discuss this issue, when Sayre mentions computers, he seems to have in mind those that are connected with the world *only* via human beings, such as the computer that a philosopher might have on his or

her desktop to use for word processing. But this is not the typical computer. In this day, most (as a straightforward percentage) computers are not on desktops or in windowless university or commercial computing centers. They are under the hoods of cars, in microwave ovens, in washing machines, in video cassette recorders, in refineries, and so forth. The internal states of these computers represent what they represent in virtue of causal connections with the world that are not mediated by humans. For example, one of the internal states in RAM of the computer in my Plymouth represents throttle position in virtue of a causal connection between the computer and my carburetor, effected by a suitable transducer and connecting wiring. It just is not true that, as Sayre says, the states of a computer fail to have semantic properties in the absence of a human interpretation. The state of the computer in my car represents conditions in the engine of the car in both the information-theoretic sense Sayre discusses and, I would hold, in the semantic sense Fodor relies on, and in both cases because of the physical structure of the vehicle.

IV. MENTALITY AND INTENTIONALITY

In his extensively revised 'The Primacy of the Intentional', Roderick Chisholm presents a case for regarding reference and other semantic features of language as derivative from the intentional properties of thought. This is the classic philosophical position. On this view, language is a tool for communicating thoughts and is parasitic upon thought for semantic properties. For example, what makes 'horse' refer to horses is that this symbol is used by English speakers to communicate thoughts about horses. In this century, this view has come to be regarded as naive since, for example, it presupposes that *language* is correlated with a totally unobservable phenomenon, *thought*. Behaviorism – and especially the influential criticism of this view by Wittgenstein – led to a new orthodoxy, in the second half of this century. On this view, the semantic properties of language are *independent* of mental representations, if there be such. In the form that Chisholm rejects, the position supposes that sentences represent *propositions* that do not change in truth value as the world changes, that believing and similar mental states are *relations to* propositions, and that the intentionality of thought is *dependent upon* the reference of language.

Chisholm does not criticize the view he rejects except in passing: he sets out an alternative account. Using an undefined notion of the attribution of a property to something, Chisholm analyzes all belief as a relation between a

believer and properties. For example, that person x believes y to be F , where y is something other than x himself, is analyzed as ‘there is a relation R such that (a) x bears R only to y and (b) the property of bearing R to just one thing and to a thing that is F is one that x [attributes to himself].’ Chisholm goes on to give analyses of other intentional relations, including *endeavoring* and *perceiving*. The latter he treats as a special form of belief origination. Finally Chisholm treats *reference* and *meaning* of language as deriving from the thoughts a speaker intends to convey by using language. Thus he defends “the primacy of the intentional” by providing an analysis founded on the primitive of the non-linguistic attribution of a property to something.

Chisholm makes *self-awareness* a condition of belief. If Chisholm is correct, no system has beliefs about the world unless it has beliefs about itself and the relations it bears to the world. Since they fail to meet this condition, most AI systems, such as Schank’s script-based systems, could have no beliefs about objects in the world (*de re* beliefs). Chisholm’s position here echoes Kant’s view that *self-consciousness* is a precondition of consciousness (though the argument Kant offers is opaque).

On the other hand, from an AI perspective, certain aspects of Chisholm’s account have clear appeal because his account does not involve *propositions*. Indeed, if we can find a way to realize Chisholm’s primitive attribution of a property to something, as well as the additional primitives he uses (appearing to, for perceiving, and endeavoring, needed for speech acts), Chisholm’s analysis shows the way to produce a host of mental attributes. Although not discussed by Chisholm, the classic ‘ISA’ relation in semantic networks might be a case of property attribution. If a system can attribute a property to some object by connecting a node representing the object with a node representing a property, then the system can have beliefs.

But this seems to reveal a problem of Chisholm’s analysis from the AI perspective. We are given no hint as to how the primitive intentional states are to be realized and the most plausible accounts of how one might attribute a property to something require that one have internal representations of the thing and the property. This would vitiate Chisholm’s project of seeing all reference – including that of an internal mental language – as derivative from prior intentionality of thought. But quite apart from that project one could still take advantage of his notion of property attribution as the fundamental mental act. From the standpoint of advocates of a causal theory of meaning, Chisholm’s title may reflect a false dilemma: *both* public language *and* thought could derive their referential properties from causal connections with the objects to which they refer; there is no sense in which language *or*

thought is universally more fundamental or enjoys primacy in all cases. As perceivers *and* language users, we come to know and think about things both from perception and from receiving linguistic messages about them. Since it appears that we can think about things that we have perceived without using public language (e.g., a prelinguistic infant watching passersby) and that we can also think about things that we have never perceived by thinking in public language (e.g., electrons), there may be more than one source of intentionality.

In his 'Intentionality and its Place in Nature', John Searle provides a clear introduction to the idea of *intentionality* itself, in the course of which he gives a compelling statement of the position that intentional phenomena are part of nature, i.e., of the physical world. Intentional states are those that are *about* or *of* something, such as perception, thoughts, beliefs, and desires. Searle takes ascriptions of intentional states as being straightforwardly true, when they are true, and as true of organisms. They are biological phenomena. Thus Searle's position contrasts with many of those prominent now and earlier in this century, most notably *eliminative materialism* (which holds that there are no such phenomena) and also *supernaturalism* or *dualism* (which concurs in the reality of intentional phenomena, but holds that they are super-natural), and any form of *reductionism* (which seeks to treat ascriptions of intentional states like belief as actually about something other than internal states of any organism itself):

Both beliefs and visual experiences are *intrinsic* intentional phenomena in the minds/brains of agents. To say that they are intrinsic is just to say that the states and events really exist in the minds/brains of agents; the ascriptions of these states and events is to be taken literally, not just as a manner of speaking [i.e. eliminative materialism], nor as shorthand for a statement describing some more complex set of events and relations going on outside the agents [e.g., presumably, philosophical behaviorism]. (p. 268)

Searle argues that eliminative treatments of conscious intentional states, such as thinking about something, on the grounds that there can be no distinction here between the fact that such phenomena *seem* to exist and a contrasting reality. If I think I think, then I think. Predecessors of this line of reasoning are to be found in Augustine, in Descartes, and – more recently – employed against materialism, in Saul Kripke (1971).

Searle believes that most identity theorists – in whose number he counts materialists, physicalists, and functionalists – “end up denying the existence of intrinsically mental features of the world” (p. 273). They wrongly suppose that to admit mental events is to abandon naturalism.

Intentional states have conditions of satisfaction: the belief that *p* is true only if *p*, and the desire that *q* is satisfied only if becomes the case that *q*. These states themselves represent their own conditions of satisfaction. Furthermore, on Searle's naturalistic account, intentional mental states play a causal role:

The essential feature of intentional causation is that the intentional state itself functions causally in the production of its own conditions of satisfaction [desires] or its conditions of satisfaction function causally in its production [perception, belief]. (p. 275)

Explanations appealing to desires are thus, he holds, a form of teleological explanation. Given that intentional states are natural biological phenomena, Searle concludes that some forms of teleological explanation are clearly proper and scientific. Searle's claim that conditions of satisfaction play a causal role in the production of the intentional state, however, does not appear to be generally true: it is generally true of intentional states associated with 'success' verbs (knows, sees, etc.) but it is only with *true* beliefs and not with *belief* in general. In the case of false beliefs, something other than their conditions of satisfaction produced these beliefs.

Using the analytic framework of intentional states representing their own conditions of satisfaction, Searle attempts a neat distinction between *mere desire*, say, to have a million dollars, and *intention*, say, to earn a million dollars. The latter – intention – is self-referential in that the satisfaction of an intention requires that the intention itself be a cause of its satisfaction. This way of characterizing the distinction appears to permit and even points to a Chisholmian reduction of intention to desire: to intend that it become the case that *p* is to desire *p* and furthermore to desire that one's desire that *p* cause it to be the case that *p*. But this is not quite right. If *agency* is what underlies the desire/intention distinction, then a qualifier must be added to the causal connection – it must be of 'the right sort'. That is, if I intend to bring it about that *p*, I will not have succeeded if a fairy godmother senses my desire that *p* and then herself causes it to be the case that *p*.

Even if such an analysis can be satisfactorily completed, it is not clear that Searle is right about the desire/intention distinction. If I intend to stop referring metonymously to all Californians as 'raisins', it appears as though my intention is satisfied if I merely cease referring in this fashion, even if this change in my behavior is not caused by the intention itself but rather by something else, say, a memory lapse or an unpleasant raisin experience.

Indeed, Searle's own example (intended to motivate the desire/intention

distinction) is clearly faulty. It may be true that fortuitous acquisition of a million dollars, say in a lottery or as a gift, does not satisfy my intention to *earn* a million dollars. But this has nothing to do with the intention/desire distinction. It is the fact that I intend to *earn*, not that I *intend* to earn, that precludes a non-earned windfall from being the satisfaction of my intention.

Searle intends his account to apply to organisms. But it is interesting to note that, on his account, it appears reasonable to hold that some *artifacts* have intentional states, notably those with feedback. Searle himself has a low opinion of the mental capabilities of thermostats. But more complex cybernetic systems have explicit representations of a goal state and employ a test to determine if that goal state is satisfied. Indeed, provisions for this are a feature of all programming languages, from high level 'while-do ...' statements to lower level conditional jumps. In common uses of these statements, the program is designed to test for the satisfaction of some condition and if the condition is not satisfied, actions are performed that can affect the satisfaction of the condition. The program then loops to the test, the condition is tested again, looping on continued falsity. Explanation of the behavior of such a system involves essential reference to the condition being tested. When it is unsatisfied, that statement is causally efficacious in the production of the behavior of the system. This analysis suggests that computer scientists can feel comfortable in agreeing with Searle that intrinsic intentional states, with representations of their satisfaction conditions, really do exist and figure in scientific explanations.

The suggested analysis of intentional state that I am offering here takes Searle as offering sufficient conditions for true ascriptions of intentionality. In this case, intentional systems can be seen as a type of negative feedback system. Of note is that such an approach can provide a much deeper *realism* with regard to intrinsic intentional states than can any account that confines intentionality to organisms. Searle refers to intentional states as 'higher level' properties and states, with the implication that there are lower level descriptions of organisms at which intentionality does not appear. But if we accept *artifacts* as possibly having intentional states, we find that this disappearance does not take place. This is surprising because, in the case of artifacts, unlike neurologically complex organisms, we know their lower level descriptions. We expect them to be drawn from the vocabularies of chemistry and of physics, where intentionality does not appear. But consider a low level description of a low level system, say, a system consisting of an operating thermostatically-controlled oven described at the level of molecules and statistical mechanics. This system is such that a fall in the mean kinetic

energy of those molecules interacting with the molecules of the thermostat bulb will cause an energy transfer resulting in restoration of the mean kinetic energy of the molecules. This feedback is a causal feature of this system at any level of description and if intentional phenomena are a subset of feedback phenomena, then they are intrinsic in a very strong sense: they do not disappear at *any* lower description of the complete system.

It is also interesting to note that a feedback system that embodies a world-to-system direction of fit and system-to-world direction of causation characterizing desire (to paraphrase Searle), also necessarily embodies directions of fit and causation that Searle's analysis shows to be characteristic of cognitive states such as perception. The system itself is part of the world, of course, and so its goal states may be self-referential, as in homeostasis. For in the feedback loop characteristic of desire there is a *test* of the *truth* of a goal state description. This truth test is a determination of the representation-to-world fit, i.e., of whether the goal state description is true or is satisfied. An implication of this is that desires require beliefs. To desire that *p*, a system must be capable of believing that *p* and that not *p*, i.e., of representing that the world in fact satisfies or fails to satisfy what it had in mind.

V. EPISTEMOLOGY AND COGNITION

Since Alvin Goldman has been one of the most prominent practitioners of naturalized epistemology, it is helpful to read the overview of this project in his 'The Relation Between Epistemology and Psychology'. He distinguishes three conceptions of epistemology – descriptive, analytical, and normative – and goes on to show the important – even ineliminable – contribution that empirical psychological considerations and findings make to epistemology under each of these conceptions. *Descriptive epistemology* is concerned with the actual sources of knowledge. So it is not surprising that it embraces much of cognitive science, including computer modelling of perceptual and other cognitive processes in AI. Goldman cites W.V.O. Quine, Jean Piaget, and Donald Campbell in this tradition.

Analytic epistemology is concerned with the analysis of such concepts as knowledge, rationality, and justification. Goldman's own 'historical reliabilist' account of knowledge falls in this tradition. Goldman suggests that "conceptual analysis is a species of psychological investigation", for concepts are mental representations and just what concepts we happen to have is an empirical question.

The third conception, *normative epistemology*, is Goldman's main concern

in this paper. On the face of it, psychology is empirical while epistemology – being concerned with correct, rational, or justified belief – is not empirical but evaluative or normative. Against this presupposition, Goldman presents a reliabilist account of justified belief, noting that it is an empirical matter to discover just which methods are reliable. Psychological facts about the reliability of memory and of perception clearly bear on the reliability of belief formation and hence the justification of the resulting beliefs.

In his fourth section, Goldman considers whether logic itself can provide normative epistemic rules without any help from psychology. He argues it cannot. Logic is concerned with implication relations between propositions. This has nothing to do directly with justified belief. But it might suggest a rule justifying any extension of one's belief set to include its logical consequences. Goldman cites two problems with the normative epistemic rule that one is entitled to believe any logical implication of one's present beliefs. Perhaps an implication of a belief reveals one should in fact give up the belief rather than accept its implications. And if one does not *believe* that a certain proposition is an implication of one's belief, even though it is, one lacks justification for believing the proposition. Goldman considers amendments to the suggested rule and finds they are subject to counter-example or lead to an infinite regress. The diagnosis of the problem appears in Goldman's next section. Attempts to justify beliefs on the basis of logical properties alone fail to say anything about the *causes* of those beliefs. A correct rule must specify the cognitive processes that produce beliefs and therefore must be about psychology.

In his final section, Goldman argues that true belief is not the only epistemically valuable end. Information must be easy to use: perspicuously organized and easily retrieved. As Goldman's examples show, many diverse psychological facts about human beings bear on such matters. These considerations illustrate his main point: epistemology, as he understands it, has all of cognitive psychology within its domain.

A meticulous reader may note, however, that the concept of normative epistemology broadens considerably during the course of Goldman's discussion. Originally, we sought to explicate notions of justified belief, and looked then to reliable methods of attaining truth. But now we find that our beliefs must not only be true but memorable as well! By the end of *this* path, epistemology encompasses rhetoric, graphics design, film editing, and every other discipline that has to do with the success with which information can be conveyed and made useful. This is to go too far. The epistemologist concerned with, say, counterfactual accounts of knowledge or with evaluating

arguments for epistemological skepticism, will not recognize this turf as his discipline. It may well be, as Goldman claims, that the *sharp* line drawn by positivists between epistemology and psychology is untenable. (Note that Goldman draws as sharp a line between logic and epistemology). But there may well be the usual blurry line that exists between disciplines.

A useful distinction can be drawn between methods that reliably produce truth and techniques that improve recall. It is not part of normative scientific methodology that one must purchase meters with large and clear displays or that one must cultivate friendships with NSF grant proposal referees, although both might reliably enhance one's prospects of generating truths. Canons for truth such as 'Do not accept inconsistent beliefs' or for induction and science reflecting Mill's Methods, such as requirements for control groups, when they are not completely *a priori*, depend for their contribution to reliability on very general features of the world; no contribution from *psychology* is apparent. In this regard, it is perhaps telling to note that Goldman considers only what *entitlements* logic might contribute to epistemological rules, rather than what constraints it might impose. It should not surprise us that logic cannot tell us what to believe. But insofar as logic can tell us what can't possibly be true, it appears that it can provide, of itself, guidance to normative epistemology in the form of negative rules.

Hilary Putnam is a critic of the naturalistic approaches to Epistemology in general and of Goldman's in particular. Putnam's 'Why Reason Can't be Naturalized' was originally the second of two Howison Lectures given at U.C. Berkeley in 1981. They have in common a critique of modern empiricism, especially that inspired by Putnam's colleague Quine but also by much of Putnam's own earlier work. The paper reprinted here is concerned with recent naturalistic (materialist) approaches to epistemology. He discusses two quite different families of approaches. These are, *first*, reductionist approaches that seek to find empirically discoverable laws and identities for the subject matter of epistemology; and, *second*, an eliminative response that finds no place for traditional concerns of epistemology in a purely natural world of physical organisms. Putnam contends that both of these positions are mistaken.

Putnam looks at several accounts of what he, like Goldman, takes to be a central concern of epistemology: what makes a belief justified or rationally acceptable. In the course of his argument, Putnam discusses several versions of these various positions. And at several points in his argument he relies on a critique, which he develops elsewhere [especially in Putnam (1981)], of what he calls a metaphysically 'realist' notion of truth. The notion of truth in

question is roughly that there are mind-independent facts, where beliefs, statements, and so forth are true if they correctly represent those facts. Putnam's argument against this position is certainly ingenious, if not convincing. The overall structure of this argument seems to be that, if metaphysical realism were correct, skepticism would be true; but skepticism is incoherent, so metaphysical realism must be incorrect. (One is tempted to say that it is just plain false, but that may be to presuppose the very suspect notion of truth!) And Putnam's argument for the claim that skepticism is incoherent turns on considerations of reference. The skeptic supposes that it is possible that one might actually be a brain in a vat, but that supposition is incoherent, Putnam argues, because if one were a brain in a vat one could not refer to brains and vats. This is a highly compressed account of the argument, but it should give the reader an idea of its main steps.

Putnam accordingly objects to the first naturalistic account of reason that he considers, *evolutionary epistemology*, on the grounds that it presupposes an incorrect *realist* view of truth. Evolutionary epistemology is the view that true beliefs have survival value and that, as a result, we have evolved a capacity to discover them. Putnam has many objections to this view: it presupposes the incorrect realist view of truth, a vacuous notion of capacity, and a mistaken equation of truth and survival value. The reader should not burn his evolutionary epistemologist registration card immediately, however, without considering that reasoned replies may be made to each of these objections. In particular, Putnam appears to have a mistaken conception of evolutionary fitness (as if it were a *guarantor* of survival).

Putnam next discusses and dismisses Alvin Goldman's *reliabilist account*, which he nevertheless takes to be an improvement over the evolutionary account: a rational belief is one arrived at by a reliable method. A method is reliable if leads to truth with a (much) greater frequency than falsehood. Putnam has two objections: the by now familiar one against the realist view of truth; and the interesting claim that Goldman's account fails to distinguish being justified from merely being the product of a method that just happens to be statistically reliable. Putnam's counter-example, it seems to me, stands or falls depending on how widely or narrowly we construe *method*. If, as in his counter-example, a person accepts everything the Dalai Lama says as true just because He says it, and it turns out that the Lama is infallible, we need to know exactly how to characterize the method used by such an alethically lucky true believer. If it is, say, 'Believe everything said by the local religious authority figure, no matter what the contrary evidence', then that method is surely not generally reliable and Putnam has not refuted Goldman.

With regard to Putnam's claim that Goldman's account presupposes a metaphysical realist account of truth, two replies come to mind. One is that Putnam has never actually succeeded in showing that there is something wrong with this view of truth; the other is that he objects to a straw man.

Goldman presents an account of epistemology fully relativised to the psychological capacities of humans; and, in characterizing the methods to be endorsed by normative epistemology, he characterizes them as methods that "maximize the attainment of epistemically valued ends". Goldman holds that these ends go beyond truth (including, e.g., ease of accessibility), and a variant of Goldman's reliabilism might not even count Truth (in any metaphysically objectionable sense) as among the many ends.

Next, and much closer to his own position, Putnam considers *cultural relativism*. Putnam counts Richard Rorty and Michel Foucault as representatives of this position. The position is that rationality is relative to culture and is not absolute. Putnam counts these views as naturalistic, although they are at odds with the reductionism of a mind/brain identity theorist, because they reflect a deference to *social* scientific understanding of nature. But, says Putnam, this relativism is inconsistent. Problems become apparent when we apply the position to itself, i.e., when we take the claims of cultural relativism themselves in a culturally relative way. Let us call this problem failure to meet the *reflexivity test*. Not surprisingly, Putnam disapproves abandoning a thorough relativism in favor of one that accords a privileged place to one's own culture and judges all others by its standards, a position Putnam calls 'Cultural Imperialism'.

Putnam's most extended and sympathetic discussion is of Quine's naturalistic views. Putnam considers Quine's views under two rubrics: *positivism* and *naturalistic epistemology*. The former encompasses Quine's view that a rationally acceptable theory of the world is one that most economically correctly predicts experiences. Putnam argues that this view also fails what I have called the reflexivity test. This is not surprising: the positivist verification criterion of meaning (that a sentence is true only if its method of verification is specified) itself notoriously fails the reflexivity test.

Finally, Putnam considers Quine's views on naturalized epistemology. These Putnam finds puzzling, because it is difficult to reconcile Quine's rejection of a realist notion of truth; his rejection of foundationalist or other classical notions of justified belief; and his various claims that he does not wish to rule out the normative. In the concluding section, Putnam gives a spirited defense of retaining the normative conceptions of traditional epistemology against eliminativist approaches.

In defense of Quine (and Goldman), it might be said that both provide naturalistic criteria for justified belief: Quine in the form of 'best fit' with experience; and Goldman in reliable methods. I find Putnam's attempt to view Quine's pragmatic positivism as self-refuting a bit too hasty. Indeed, Putnam is rather disarmingly misleading here. He says of the criterion of right assertibility that he attributes to Quine, "This statement, like most philosophical statements, does not imply *any* observation-conditionals, either by itself or in conjunction with physics, chemistry, biology, etc." On the one hand, he attempts to assimilate the criterion to 'most philosophical statements', a kind of guilt by association, when in fact it is not much like the philosophical statements that are most obviously disconnected from observation and thus the target of positivist strictures (statements such as 'God loves you greatly but never shows it'). On the other hand, Putnam deftly conceals psychology and all the other social sciences in the 'etc.' at the end of his list of sciences. This is important. If the natural social *function* of assertion is to provide the pieces of the most economical predictor of experiences, then the criterion for right or acceptable assertion Putnam attributes to Quine is not so obviously the self-refuting dangler cut off from scientific confirmation that Putnam asserts it is. I do not believe that Quine's account is actually correct, but I believe Putnam fails in his attempt to show it is so defective as to be self-refuting.

Putnam's own positive views are not stated here, and I find them somewhat exclusive when they are stated, as in *Reason, Truth and History*. This is a cliff-hanger in which we read white-knuckled, wondering how the author/hero is going to escape the dreadful fate that awaits all other relativists. Somehow, in all the action, the scene goes a bit out of focus and when it returns we merely observe that our author/hero has somehow miraculously survived and has gone on to slay new conceptual monsters that ravage the philosophical village (the "hangups that have marred so much recent philosophy", as he puts it). But exactly how all this has been done is never completely clear.

VI. THE MENTAL AND THE PHYSICAL

Patricia Kitcher distinguishes two versions of mind/brain identity theory and defends them against the 'absent qualia' argument. This latter is the argument that any physiological state an identity theorist might claim is identical with a certain psychological state could occur in the absence the subjective qualities of the psychological state. Thus, C-fiber firings could occur without there

being any subjective pain, we can easily imagine it so – and hence *C*-fiber firings cannot be just one and the same thing as pain. This argument appeared in the early 1970s, in Saul Kripke's 'Naming and Necessity' and Ned Block and Jerry Fodor's 'What Psychological States are Not'. But in fact in my view it merely reflects a rediscovery of a central argument of Descartes' together with the thesis of the necessity of identity that underlies both Cartesian and Kripkean arguments against mind/brain identity.

Against Kripke's argument, Kitcher notes that one must carefully distinguish *epistemic possibility* from *metaphysical possibility* – it may be possible, *for all I know* (epistemic possibility) that *x* occurs without *y*, but it may not actually be possible (metaphysical possibility). The distinction is easy to see in matters mathematical: it is possible, *for all I know*, that 2457 times 39 is 95723, or 95873, or any of many other numbers, but the product is either necessarily 95723 or necessarily something else (it is not metaphysically possible for the product to be anything other than what it is). Thus while it may *seem* to me that *C*-fiber firings may occur without there being any sensation of pain, this seeming is merely *epistemic* possibility, and identity theorists could simply deny that there is any correlative *metaphysical* possibility of non-identity. Much the same reply can be made, I think, to Descartes' argument (for 'the REAL distinction' between mind and body) that, since we can imagine or conceive the mind existing without the body, they cannot be identical.

The two versions of identity theory distinguished by Kitcher are the 'psychological theory' and the 'physiological theory'. The former holds that mental terms refer to psychological states and that a description of the underlying physiology is not perspicuous nor can it play a useful role in explanations involving mental (= psychological) states. The physiological theory, on the other hand, holds that mental terms actually refer to physiological states and denies such pessimism about the perspicuity of physiological explanation. Kitcher impartially defends both in this paper.

In reply to the objection to the physiological identity theory – that it is implausible to hold that people really are referring to physical states when they refer to mental states (since most people don't know anything about neurophysiological states) – Kitcher relies on views of reference developed by Kripke, Putnam and others. Put simply, the new theory of reference has the implication that one need not know anything about something in order to refer to it. Views of reference to the contrary – which held sway during the previous half century – confuse epistemology and semantics. Kitcher notes that most adults can recognize and refer to *gin* without being able to provide

physical criteria for distinguishing gin from other liquids. The same can be true of brain states: we can recognize them by their subjective qualities without knowing precisely how we do it.

There are two possible replies to the absent qualia argument that Kitcher considers on behalf of the psychological theory. She seems to think they are equally appealing. The second, however, seems much more promising to me. The *first* is for the psychological theorist (e.g., a functionalist) to concede to the critic that the functional account is indeed incomplete and to provide a small but crucial place for physiology. Obviously this is to abandon functionalism in the crunch (see White (1986) for extended criticism of this approach). Furthermore, for this strategy of appealing to physiology to succeed, reliance must be made on a reply to the absent qualia argument directed – as by Kripke – against the physiological identity theory. But then, why not just use this strategy in defense of the psychological theory itself? Indeed, this is the *second* strategy. And from the standpoint of the functionalist, this line of defense has the advantage of being much more perspicuous than the claim of the physiological theorist. It seems – at least to me – that when the functional role of pain is fully characterized, it will be much easier to see that a state that plays the exact functional role of pain must necessarily have the subjective quality of pain than it will ever be possible to see, even given a complete neurophysiological description of the brain, that certain neuron firings must have the quality of pain.

William Bechtel, in ‘The Functional Architecture of Mind’, represents the anti-reductionist position. He accepts Fodor’s and Putnam’s objections to type/type identity between psychological and physiological states and processes; and with the rejection of those identities goes the hope of reductions of psychology to physiology. Thus he sides with functionalists such as Putnam, who believe that there can be many quite distinct physical realizations of the same psychological state; and with those such as Fodor, who argue that the same physical state may realize quite different psychological states, depending on contextual factors. But functionalism is not an abandonment of materialism. A functionalist can suppose that each *individual mental event* (an individual instance of a type is called a ‘token’) is identical with some *individual physical event* or other (which makes precise the notion of ‘instantiation’ used above in characterizing the objections to classical identity theory). That is, the functionalist can embrace a ‘token-token’ identity of mental and physical. While Bechtel does not deny the truth of token-token identities, he believes they do not provide the most “useful handle on the relation of cognitive science to neuroscience” and he proposes a composi-

tional account of the relation.

The key to Bechtel's positive suggestion is the notion of the functional architecture of a system, such as the central nervous system. Bechtel attributes this notion to Pylyshyn but, as Bechtel uses it, it has many antecedents, including Fodor's notion of modularity, among others. The functional architecture of a system can be discovered by 'cognitive impenetrability'. This in turn is related to the hoary *nature/nurture* distinction: some processes have outcomes that can be altered by information (e.g., learning); some do not. An important example would be optical illusions. Even when subjects are informed, or discover, that the lines in a Mueller-Lyer figure are of equal length, the illusion that they are not persists. This reveals a level of process in visual perception that is quite unaffected by the information that the lines are equal in length: a functional module accepts input from the eye and presents information regarding relative length as output to higher cognitive modules in the system. Fortunately, of course, the information from the visual module can be successfully evaluated as misleading at the higher level. But the entire system, the mind, that produces judgments of relative length, is composed of functional parts.

Bechtel is therefore concerned about developing a proper understanding of the role of neurophysiology. In his second footnote, he discusses Pylyshyn's own pessimistic assessment of the prospect for neurophysiology as a source of *discovery* of the functional architecture of the mind. But Bechtel believes that those neurophysiologists who study interactions between subsystems in the central nervous system are in fact studying the functional architecture of the mind, for they are studying information processing:

My proposal, then, is to think of information processing as a higher level process involving the interaction of the individual neural processes that fall in the special domain of neuroscience. Those neuroscientists who focus on the interactions of neural components are studying information processing. ... They are working on the same side of the functional architecture as cognitive scientists who try to model how the brain manipulates symbols. (p. 371-2)

Many functionalists will find little to disagree with in what Bechtel proposes and will not find his conceptual framework incompatible with functionalism. While functionalism is compatible with token-token identification of mental events with brain events, a functionalist need not believe that *all* or even that *any* such identities are true [see Cole and Foelber (1984), 'Contingent Materialism' for a more complete explanation of how this can be so]. A dualist could suppose that the mind has a functional architecture. And those functionalists who do suppose that all mental events are physical might

readily concede that these identities are not especially illuminating, any more than knowing that a student grade-roster contained in a computer system is instantiated by the electrical state of a particular RAM chip. In particular, the electrical facts tell us nothing about how to get the roster sorted alphabetically.

But it is less clear how to appraise Bechtel's critical remarks. These identities have their place. If we want to know how a computer is able to process information, at some level our discussion must turn to transistors and voltages. Bechtel suggests that his compositional account is a real alternative to *all* identity theories, but it may be a distinction without a difference. To say that the U.S. is composed of 50 states or that the human body is composed of various organ systems is to make identity statements. A statement identifying a system with a certain arrangement of subsystems is at once an identity, reductionistic, and compositional. In any case, the most interesting claims in Bechtel's piece concern the role of neurophysiology in cognitive science and, in particular, his suggestion that some neurophysiologists are studying the brain as an information processing system. Thus, the apparent gulf between neurophysiology and psychology is bridged.

VII. EPILOGUE: CONFLICTING CONCEPTIONS

The syntactic or computational theory is essentially a causal theory. Such a causal theory attempts to account for internal mental processes and thought (inference, association, hypothesis formation, categorization, pattern recognition, etc.) on the basis of syntactic operations. It is appealing, of course, because syntactic operations are well understood in that we know very well how they can be performed by a physical system. Furthermore, as Fodor has noted, it seems that any given modular process within a mental system only has access to syntactic features of other states of the system. Information, as it is passed from one module to another, must be represented by syntactic features of the signal.

James Fetzer argues, however, that the computational, or syntactic, approach to understanding minds pays much too high a price for its putative advantages. Fetzer views computational, representational and dispositional theories of mind as parallels of syntactic, semantic and pragmatic accounts of language. Just as syntactic and semantic treatments are necessarily incomplete accounts of language – even of meaning as conveyed by *speakers* – so too are purely computational or representational approaches to the mind inadequate.

Fetzer argues that a computational account of mentality is the most clearly inadequate. Computational theories are committed, says Fetzer, to one or another, if not both, of the theses that mental tokens have the same content if and only if they have the same form. But there are clear counter-examples to both conditionals. Attempts to salvage the computational approach by appeals to context, which must of course be syntactically specified, are of no avail because ambiguity – that is, multiple possible contents – remains.

Representational accounts are a genuine improvement over computational. However, because they disregard pragmatics, the role of language in the behavior of the organism, these accounts of mentality are inadequate as well. Fetzer discusses Fodor's thesis in *The Language of Thought* as the principle representative of representational views. Fetzer draws an analogy between Fodor's hypothesis of an innate language of mental representations and Plato's doctrine of knowledge as recollection, suggesting that the doctrines are approximately equally plausible. The primary problems with the language of thought thesis are that it presupposes an inadequate theory of meaning (meaning as truth conditions) and that it ultimately fails to account for the semantic properties of primitive language.

A dispositional account can succeed where its rivals fail, Fetzer argues. He draws on C.S. Peirce and offers an account of meaning that is functionalist, but which displays the behaviorist pedigree of functionalism more than other such theories. On this account, the meaning of a particular belief, for example, is "... the totality of tendencies that the system would possess in the presence of that belief ...". (p. 396) The account is functionalist rather than behaviorist because these tendencies include those to internal state transitions as well as to public behavior. The theory is pragmatic, making the role of the mental state in the life of the organism relevant to meaning and mentality. Yet the account breaks with Wittgenstein and other pragmatic accounts of meaning, such as that of H.P. Grice, in that it requires only a single symbol user and does not suppose that meaning and mentality are dependent upon the characteristics of, or can only occur within, a community of language users. It therefore supports the possibility of private languages and private language users.

Fetzer's discussion is clearly interesting and important. Does he succeed in showing that computational and representational accounts are false? In any domain as complex as this – with such subtleties as distinguish variant emerging theories – avoidance of strawman argumentation is difficult. Fetzer clearly takes his dispositional account to be a rival of the other accounts. Yet, for several reasons, it is not clear to me that his account is incompatible with the others. There may be much that a computationalist might agree with in

Fetzer's account of content of concepts. For one thing, Fetzer's account, with its emphasis on symbol *users*, appears to apply primarily to natural language symbols (his examples in argument are English sentences). It is less clear how this account extends to cover the mental abilities of lower animals and the prelinguistic abilities of higher animals. It is difficult to see how Peirce's account is of much help here, although Fetzer offers suggestions directed toward that end in his "Signs and Minds".

Second, Fetzer's account explicitly requires that the system be capable of being *conscious* of the sign: "For any semiotic system ... a sign, *S*, stands for something *x* for that system rather than for something else *y* if, and only if, the strength of the tendencies for that system to manifest behavior of some specific kind *when conscious of S* – no matter whether publicly displayed or not – differs in at least one context ..." (p. 396, emphasis added). But much, perhaps the bulk, of work in cognitive psychology, linguistics, and artificial intelligence concerns states of which the system is not conscious. Again, on its face the account seems more amenable to linguistic abilities of humans than to a general account of mentality. Fetzer indicates in note 2 that he takes consciousness of a symbol merely to be the capability to exercise an ability to use a symbol. Much thus depends upon how 'use' is to be construed here. Since Fetzer wants to include some lower animals as symbol users, presumably his sense of 'use' is not a traditional sense in philosophy of language which requires conscious intentions. But as his sense of consciousness becomes inclusive of the minds of lower animals, it becomes less clear how it excludes the syntactic engines of the computationalists – a robot such as Shakey at Stanford Research Institute in the 1970s used symbols, in *some* sense of 'use', to represent its world, and was such that "the strength of tendencies to manifest behavior of some specific kind" when in syntactic state *S* differed from those when it was not.

Third, some computationalists, such as Stich, simply disavow any interest in meaning. Such a computationalist might *agree* with everything Fetzer says about meaning, because he/she does not regard meaning as a problem being addressed by the computational account.

This brings us to what I take to be the main worry about Fetzer's overall argument [a more complete presentation of his position may be found in Fetzer (1990)]. A computationalist who *is* interested in meaning and *is* interested in global aspects of the organism and its behavior can agree with much of what Fetzer has to say about the roles of "other beliefs, motives, ethics, abilities, capabilities, and opportunities" in meaning and behavior. But Fetzer's analysis seems to stop at a level of explanation that is not the primary concern of the computational theory. For the computationalist can

ask the simple but obvious question: what accounts for ‘the tendencies to behave’ that are the bedrock of his dispositional analysis? For example, suppose x has a ‘tendency’ to infer from the belief that Jim and Mary went to the store that Jim went to the store. How is that inference made? What accounts for that tendency to infer? The computationalist is, in part, attempting to answer questions such as these. Such functional tendencies are problems for the computationalist, but answers for the dispositionalist.

Fetzer is undoubtedly correct, however, in holding that a computational account of itself is inadequate as an account of the content of mental states. Some computationalists, such as Stich, attempt to make a virtue of this by dissolving the problem, arguing that the whole notion of belief is ultimately incoherent. (Much of Stich’s argument, suspiciously, rests on scenarios where we allegedly are at a loss as to what belief can be ascribed, cf. *From Folk Psychology to Cognitive Science*, pp. 137-148). But a *causal* theory, of which a computational account of mental states is but a part, might be able to account for semantic properties, the content or intentional or representational properties of mental states, through causal relations as well. Causal theories of intentionality involve states outside the system. (After all, semantics is the relation of symbol to world). What makes a particular mental state M represent a particular state of the world W appears to be a reliable causal relation between M and W , mediated by sensory processes of the system.

Whether a causal theory will ultimately resolve all the problems in computational and representational accounts pointed to by Dretske, Sayre, Searle, and Fetzer remains unclear. But I believe this is the direction to which we should look for solutions. The papers that follow provide ample indications of the conceptual problems confronting the development of an adequate theory of mind, together with many capable sketches of alternative possible solutions. They are valuable resources for future research in the theory of mind and in artificial intelligence.

NOTES

¹ John McCarthy: (1988), ‘Mathematical Logic in Artificial Intelligence’, *Daedalus* 117 (1, Winter), p. 305.

² Thus in looking at the problem this way we practice what might be called ‘academic ascent’, paralleling Quine’s notion of semantic ascent.

³ I read Demopoulos’ account in ‘One some Fundamental Distinctions of Computationalism’ as having the same consequence. [Demopoulos (1987), p. 89]

PROLOGUE

WHAT IS MIND?

CURRENT ISSUES IN THE PHILOSOPHY OF MIND

The philosophy of mind is one of the most active fields in philosophy today, and it has changed so drastically in the last twenty years, that many of the traditionally central topics and theories have been transformed almost beyond recognition, and new concerns now loom that have no clear ancestors in the old tradition. An assessment of current work requires an understanding of recently evolved assumptions about the burdens and goals of the field, which can best be provided by a brief history of the shifts of outlook in recent years.¹

I. INVESTIGATING THE LANGUAGE OF MIND

The new era in philosophy of mind can be dated from the publication in 1949 of Gilbert Ryle's *The Concept of Mind* [93]. In that book Ryle argued that the philosophy of mind rested on a colossal error, a "category mistake" that had in effect given birth to a whole field of investigation – the philosophy of mind – where none ought to be; the questions composing the inquiry were so radically misconceived that straightforward attempts to answer them ineluctably led to nonsense. Before Ryle there had been *theories* of mind – such redoubtable "isms" as idealism, materialism, neutral monism, epiphenomenalism, interactionism – contending in an arena of shared assumptions about the nature of the problem defining the field: the mind-body problem. Ryle suggested there was no such problem at all, but only a confusion bred in an injudicious and insensitive use – or abuse – of the ordinary language we use in everyday life to aver the familiar facts of mentality that comprise the data for any investigation or science of mind.

A careful *analysis* of the ways of common talk about the mind would dissipate the confusions, dissolve the problems, and thus render otiose both dualism and its negation, monism, and among varieties of monism, both idealism and its opposite, materialism, in short all the rival metaphysical views that had been the chief product of the field. Ryle's work was the major entry of what came to be called *ordinary language philosophy* into the philosophy of mind, and its influence should be measured not by a census of converts (philosophies seldom display their influence by attracting

proponents, but rather by remaining controversial for long periods of time), but by the fact that for more than a decade after the appearance of *The Concept of Mind, theories* of mind were unfashionable to the point of extinction. Theories were held to be the creations of those who had failed to see that the problems – at least the problems philosophers were equipped to address – arose from mistaken and naive assumptions about the way mind-words worked in the language. Theory-construction was replaced by the much more cautious and modest activity of “conceptual analysis”: the delicate and persistent, if informal and unsystematic, canvassing of the idioms of ordinary language, the collective product of which is a broad and still largely unsystematized array of acutely observed distinctions and nuances, adduced in the course of making usually quite small points about various mental concepts. At its best, in the work of Ryle, Wittgenstein, Austin and Anscombe [5, 8, 10, 93, 117], this method uncovered deep conceptual issues that still shape current thinking and will continue to do so. At its worst, like the worst in any field, it produced mountains of trivia, but in the middle there was a good deal of very clever and useful work, whose point was seldom to solve problems and almost never to advance general theories, but typically to alert the incautious to the existence of more problems and distinctions than one would have expected [43, 63, 73, 94, 113, 114, 115, 116]. There seemed in those days to be very little in the way of a substantial generalization that could be defensibly advanced.

The basic tactic then was “semantic ascent” [85]; when one runs into perplexities when talking about things (in this case, minds, sensations, thoughts and the like) it often helps to shift one’s focus and talk about how to talk about those things, about “what one would ordinarily say” under various circumstances, or (if one is not enthralled by *ordinary language*) what one *ought to* say under various circumstances. There is no denying the value of the tactic, as the great work in the field amply demonstrates, but contrary to the creed of many at the time, it has not turned out that *all* the problems in philosophy of mind evaporate under linguistic analysis, and it is fair to say that a great deal of the researches into ordinary idiom failed to produce anything more important or enlightening than an intense appreciation for the subtlety of English expression. Moreover, although semantic ascent excuses one from expounding or defending an “ism”, it does not permit one to operate innocent of assumptions, assumptions which ultimately implicate one in something rather like a theory. Typically (for Ryle, Malcolm, Anscombe and many others) the tacit theory was “logical behaviorism”, the view that the truth of ascriptions of mental states and events implies and is implied by the

truth of various statements purely about behavior. There were dissenters from this view: Strawson [108] for instance apparently committed himself to a cryptic revival of the *double-aspect theory* (a *person* is not to be analyzed into a body plus a mind, but is nevertheless a proper subject of both mental and physical attributes) and on Shaffer's analysis [99] ordinary language was held to incorporate dualism.

This tendency to evince theory in the course of conceptual analysis had the effect, at least in its most virulent forms, of squandering the useful indirectness of semantic ascent altogether, and bodies of doctrine emerged that looked suspiciously like the old metaphysical theories about the *things* (minds, sensations, thoughts, ...), though generated from considerations scrupulously restricted to *words* ("mind", "sensation", "thought", ...). This was not a happy development. Perhaps you can learn all about the *concept* of a horse by studying the way we ordinarily use the word "horse", and no doubt you can learn a great deal about horses from studying the concept of a horse (since much of what people *think* is true of horses is embodied somehow in our concept), but in the end there are some left-over facts about horses of non-negligible interest and even puzzlement that can be discovered only by looking at a horse or two, or at least by reading the works of those who have taken the trouble to do this.

II. PHILOSOPHY OF MIND NATURALIZED

Ordinary language philosophy of mind has now played itself out to the point where it can be comfortably viewed as a historical phenomenon. As an essentially critical and reactive discipline, it was bound to die of its own successes when it had run out of important errors and confusions to diagnose, while its infirmities became more apparent as it descended into trivia. Although its most characteristic doctrines and methods have been widely rejected or abandoned, its contributions to current thinking are positive and pervasive. Most important, the new way with words really did destroy the traditional way of composing a philosophical theory of mind. The traditional theorists were guilty, as charged, of making aprioristic generalizations, which were nothing if not the products of (ill-considered and un-self-conscious) conceptual analysis, mixing these with a handful of casual introspections and observations about normal people's experiences and powers, and promoting the mixture to the status of metaphysical verities about the essences of things mental. If there were to be theories of mind at all, they were not to be produced by the old armchair methods, so the philosopher of mind had three

choices: abandon philosophy and pursue empirical theories in the domain of psychology or brain science, abandon theory and settle for the modest illuminations and confusion-cures of purely linguistic analysis, or become a sort of meta-theorist, a conceptual critic of the empirical theories advanced by the relevant sciences. It is this last conception of the enterprise, where it is seen as a branch of philosophy of science, that dominates the best work in the field today. Its most salient difference from both the traditional theorizing and the ordinary language approach is its interest in the theories and data of psychology, the brain sciences, artificial intelligence and linguistics.

In 1932, H.H. Price's classic work, *Perception* [77], contained a succinct apology for the philosopher's ignorance of science: our grounds for believing the physiological accounts of perception "are derived from observation, and mainly if not entirely from visual observation." But the reliability of observation is just what is at issue for the epistemologist, and "since the premises of Physiology are among the propositions into whose validity we are inquiring, it is hardly likely that its conclusions will assist us." So long as one is engaged in a Cartesian attempt to justify all knowledge from scratch, from whatever minimal foundation can be protected from systematic skepticism, this familiar rationale can be maintained (though it provides no good reason not to *peek* at physiology), but such foundationalism in epistemology and philosophy generally is now on the wane, replaced by a "naturalistic" attitude (not a theory) that assumes from the outset that by and large our quotidian beliefs are true and warranted, that epistemology can learn from psychology, and that the best way to derive the canons of justification is to see how good science is done. (The attitude is well expressed by Quine, one of its most influential promoters, in 'Epistemology Naturalized' [86].)

The danger of this drift might seem to be that it leaves philosophy with no standpoint from which to launch truly radical critiques of current science, but this is probably a misconceived worry, for the history of science suggests that revolutions in scientific thought must be internally bred. Still, the claim that the new naturalism is a capitulation to the excessive prestige of modern science has something to be said for it, though it should not be forgotten that philosophy's current friendship with science is not a novelty. The great philosophy of the 17th Century, for instance, was in intimate communication with the contemporaneous birth of modern science and contributed as much to that infancy as it gained in return. In any event, philosophers have discovered a vein that will be mined, very probably to the mutual enlightenment of science and philosophy.

While this emergence from the intellectual isolationism of the recent past

is thus a logical development out of the best in the linguistic analysis tradition, it nevertheless required the rejection of a troika of doctrines central to that tradition: verificationism, logical behaviorism, and what might be called conceptual conservatism. In its most exigent form, verificationism is the doctrine that the method of verifying the application of a term just *is* its meaning: in its milder forms verificationism maintains that claims that cannot in principle be verified are senseless. This surviving brainchild of logical positivism is a plausible enough doctrine until one gets severe or doctrinaire about what is to count as verifiability, as typically happens. Consider, for instance, this simplified statement of the notorious “problem of other minds”. How do I know that other people have minds? I cannot directly see or otherwise sense their minds (as I can introspect my own); all I have for data are observed facts about their behavior. Perhaps their behavior is good inductive *evidence* for the existence of their minds. But it could not be, for in order to establish that it was good evidence, we would have to have *confirmed* cases of the co-occurrence of such behavior with other minds, and this requires, *per impossibile*, an independent method of verifying the existence at those times of those other minds. As the slogan had it, sometimes can be a *symptom* of *x* only if something else is the *criterion* of *x* [1, 54]. Criteria were thought to be not (merely) *empirically reliable* but rather *logically sufficient* or at least *decisive* or “certainty-providing” indicators of whatever they were criteria for. Then, since an appeal to ordinary language shows that the claim that there are other minds is not senseless (the man in the street knows full well there are other minds), it must be verifiable, and since the only evidence by which to verify it is behavioral, and since symptomatic evidence is logically dependent on criterial evidence, *there must be purely behavioral criteria for all (meaningful) claims about other minds*. Thus is logical behaviorism born of verificationism. But now suppose some mental item, say *pain*, does have purely behavioral criteria. That means that the assertion that someone is in pain is really (logically equivalent to) a statement about that person’s behavior or dispositions to behave. But no statement about inner physiological happenings could be logically equivalent to a statement just about a person’s overt behavior, so the truths of physiology, whatever they turn out to be, are *irrelevant* – except symptomatically – to the truth of claims about pain. Since the concept of pain has behavioral criteria, it cannot also have physiological criteria. Were scientists to propose physiological criteria for pain, they would be “proposing a new concept” of pain, and anything they told us about *their* sort of pain would not be about our *ordinary* concept of pain at all. Science cannot revise or improve on

ordinary concepts, but is bound to abide by the criteria of use enshrined in ordinary language, on pain of either changing the topic or talking nonsense. Thus is conceptual conservatism born out of logical behaviorism.

The development of this line of thought in the literature was immeasurably more subtle, guarded, and attenuated by provisos and acknowledgements than the sketch of it given here (see, e.g., [54]) and of course there was much in it that was true, but it lent support to a dubious claim; if psychologists and neurophysiologists thought they could study the mind, they were wrong; the study of mind was the study of (ordinary) mental *concepts*, and since these had ordinary behavioral criteria of application, once these criteria had been adumbrated by philosophers, there was nothing left to do. This message was intolerable even to many of the adherents of the method of ordinary language analysis that had led to this embarrassing result. It was the work of many hands to dismantle the edifice of argument and assumption that led to this impasse but the whole processs is graphically epitomized by Putnam's classic attack [79] on Malcolm [59, 60], by Fodor's polemics against Ryle [38, 39] and by Chihara and Fodor [16].

III. THE IDENTITY THEORY AND ITS DESCENDANTS

The first proclaimed alternative to logical behaviorism to draw serious attention was the identity theory of mind: minds are brains, and the contents of minds – pains, thoughts, sensations, and the like – just *are* (identical with) various happenings, processes and states of our brains. The early papers supporting the identity theory, by Place, Feigl, Smart, and Armstrong [7, 31, 32, 76, 101, 103], had the flavor of manifestos, and their point was to secure as directly as possible what was deemed to be the requisite conceptual foundation for a purely physicalistic or materialistic science of the mind, a bulwark against both the impertinent dismissals of the logical behaviorists and the metaphysical excesses of dualistic alternatives.

(Since it is widely granted these days that dualism is not a serious view to contend with, but rather a cliff over which to push one's opponents, a capsule “refutation” of dualism, to alert if not convince the uninitiated, is perhaps in order. Suppose, with the dualists, that there are non-physical effects (or accompaniments) of brain events. Then either the occurrence of these effects has itself *no effect whatsoever* on subsequent events in the brain (and hence behavior) of the person (epiphenomenalism), or it does (interactionistic or Cartesian dualism). In the former case the postulation of the non-physical effects is utterly idle, for *ex hypothesi* were the effects to cease to occur

(other things remaining the same), people would go right on making the same sorts of introspective claims, avowing their pains, and taking as much aspirin as ever. Even more vividly, were a person's epiphenomena to be gradually delayed until they ran, say, ten years behind her physical life, she and we could never discover this! In the latter case of interactionistic dualism, since the occurrence of *non-physical* events (events having temporal location and presumably particular-person dependency but lacking spatial location and mass-energy) would be required to trigger unproblematically physical events in the brain, the conservation laws of physics would be violated. Either way, one pays an exorbitant price for dualism.)

The identity theory was to be an empirical theory, conceptually outlined by philosophy but with the details filled in by science, and its ontology was typically presumed to include only scientifically well-credentialed entities – no *élan vital*, no psi forces, no ectoplasm, only brain cells and their biochemistry and physics. The identity theory's defining claim, the claim that mental events are not merely parallel to, coincident with, caused by, or accompaniments of brain events, but *are* (strictly identical with) brain events, divides people in a curious fashion. To some people it seems obviously true (though it may take a little fussing with details to get it properly expressed), and to others it seems just as obviously false. The former tend to view all attempts to resist the identity theory as motivated by an irrational fear of the advance of the physical sciences, a kind of humanistic hylephobia, while the latter tend to dismiss identity-theorists as blinded by misplaced science-worship to the manifest preposterousness of the identity claim.

This antagonism has created a very large literature over the last fifteen years – much of the best of it [7, 20, 23, 31, 32, 34, 52, 65, 76, 88, 101, 103, 110], is anthologized in Borst [15] – and out of it has emerged a panoply of sophistications that leave the original bluff identity theory far behind while advancing basically unrevised its basic project of providing a conceptual pedigree for the physical sciences of the mind. The difficulties encountered by the identity theory can be divided without major loss and distortion into three basic areas: problems arising from Leibniz's law, problems about generalization, and abstract logical puzzles about the identity relation.

It is Leibniz's law that makes identity a stronger logical relation than mere similarity, co-occurrence or equivalence. It states that “*x* is identical with *y*” entails that *whatever* is true of the thing denoted by “*x*” is true of the thing denoted by “*y*” and vice-versa. The principle is unassailable, since in the case of any true identity “*x*” and “*y*” will denote the very same thing, and whatever is true of that thing is true of it, whatever it is called. But now suppose some

thought of mine is witty, or profound, or obscene; the identity theory must then claim that some brain process or event (the brain process or event identical with that thought) is witty, profound or obscene, and at least at first glance brain processes don't seem to be the type of thing that could be witty or profound or obscene, any more than they could be the square root of 7 or loyal or capitalistic. One reply to this objection, variously expressed and defended, is: take another glance, and you will see that a brain process *can* be witty or profound or obscene in just those cases where it happens to be a thought. Not all events in the brain are thoughts – some are just metabolic events, for instance, and they cannot be witty, but they are not the only brain events. Of course the success of this position depends heavily on having an account of what it could be about a brain process or event that made it a thought (and one thought rather than another), and as we shall see, there are important problems in this area. (Another plausible reply to the objection distinguishes the thought as *event* from the thought as *content* or *proposition*, and claims that such features as wittiness properly apply to the content, not the event, but again, this position is no stronger than one's theory of the individuation of events by their content.)

There are in any case apparently harder problems raised by Leibniz's law [55]. Suppose I am subjected to visual stimulation that subsequently produces in me a round, orange after-image. There is no round, orange image on my retina of course. Are we to suppose that there is a round, orange brain event or brain state that is identical with my after-image? Nothing that is not round and orange can be identical with something that is round and orange, so either my after-image is identical with some round, orange brain-state (which no one believes, I trust), or my after-image is not round and orange (could we claim it just seemed to be?), or my after-image is something (mental) other than – in addition to – any brain event and hence the identity theory is false, or there simply are no such things as round, orange after-images. It is no doubt tempting to anyone unfamiliar or unimpressed with the conceptual horrors of dualism to abandon the identity theory at this point, and admit after-images and their ilk to his ontology as extra non-physical, epiphenomenal by-products of brain activity, but this is just what the identity theorist refuses to do [19]. The favored step instead is the last one: to deny that there are, strictly speaking, any such things as after-images at all. What there are, we are told, are havings-of-after-images, or experiences-of-after-images, and *these* are neither round nor orange. The difference between experiencing an orange after-image and experiencing a green after-image is not that the former experiencing is orange while the latter is green [19, 23],

93, 101, 103, 104]. This move calls into question what might be called the normal semantics of ordinary mind-talk. We speak, casually and ordinarily, using words like “pain,” “image,” “belief,” “brainstorm,” “hunch,” etc., as if these words were unproblematic referring expressions denoting perfectly real items in our minds (whatever they are). But perhaps our ordinary talk embodies a fossilized myth-theory, and once science teaches us what is really happening in our heads, we will abandon the search for referents for our ordinary mind-terms, just as we have abandoned our search for mermaids, witches, and demons. We talk as if there really were such things as after-images, and so there seem to be; we also talk as if the sun really rose in the east, and so it seems to, but science can render the former mode of expression, like the latter, metaphorical. This line of reasoning can lead in the extreme to a variety of positions often gathered under the rubric of the “disappearance form” of the identity theory: science will not *discover the identities* of the problematic mental items, but rather those items will disappear as candidates for identification as a more sophisticated scientific picture supercedes the old [23, 34, 88, 109].

The second set of problems with the identity theory concerns generalization. These problems arise because the normal role of identity claims in theories is to permit generalization. (If this cloud is identical with a collection of water droplets, then so perhaps are the others; if this gene is a DNA molecule, the tempting hypothesis to test is that all genes are.) But it is far from clear that the identity theory could – or should – provide us with *any* generalization beyond its umbrella claim that every mental item is identical with some brain item or other. Suppose Mary thinks about *pi* at noon, and the identity theorist claims that Mary’s thought is identical with her noontime brain process *p* (having defining physical features *F, G, H, ...*). It is not remotely plausible to suppose that every *thought about pi* is a brainprocess with features *F, G, H, ...*, if only because there is no reason to suppose intelligent creatures elsewhere in the universe would need to share our neurophysiology or even our biochemistry in order to think about *pi* [81]. It is not even plausible that every *human thought about pi*, or even every *thought of Mary’s about pi* is identical with a brain process falling into a class specifiable solely in terms of the physical features of the members. It would seem to be a burden of any theory of the mind that is tell us what it is about thoughts that makes them thoughts, and what it is about thoughts about *pi* that makes them thoughts about *pi*, and it does not appear that the general features we are looking for are physical features. The *weak* reply that the identity theorist can make is that all he needs to claim in order to avoid

dualism is that each *particular* mental event (each “token”) is some brain event or other (no mental event is a *non*-physical *non*-brain event). Thus we distinguish a “token” identity theory from a “type” identity theory, and abandon the latter in philosophy of mind. Perhaps no one today supposes that *types* of mental items can be distinguished directly by purely physical features, but almost no one any longer supposes this was a reasonable goal of physicalism [20]. The *strong* reply goes beyond the avowal of a token identity theory and claims that the sought-for distinguishing marks of the types of mental items are definable in terms of *causal* roles filled [7, 52], or in terms of the *logical* states of the abstract Turing machine “realized” by a human being’s nervous system [78, 80, 81, 82, 83], or in terms of the *functional* roles filled [37, 38, 39]. As the differences between these three views have been sorted out, and deficiencies noted and corrected, a single widely shared view has emerged called *functionalism*: mental states are functional states, that is, states individuated by their functional role within the whole system. To say that a particular belief, or pain, for instance, is a particular functional state, is to say that anything, regardless of its composition, chemistry, shape, or other physical feature, that fulfilled the same functional role in a functionally equivalent system would be the same belief, or pain, and nothing that fulfilled such a functional role could fail to be such a belief, or such a pain. Functionalism has become the dominant doctrine in philosophy of mind today (that is, it is the only theory being widely criticized and defended in the journals), and hence will receive more detailed discussion below. It should be already clear that this sort of functionalism has little to do with the brand of functionalism encountered in sociology or anthropology.

The third area of investigation initiated by the identity theory concerns the logic of the identity relation itself. It was initially supposed by Smart and other early defenders of the identity theory that the concept of identity was perfectly safe and well understood. It was a common tactic to elucidate the identity theory of mind by drawing analogies to presumably innocent and familiar identities encountered in less puzzling quarters, such as the identity of lightning bolts with electrical discharges in the air, the Morning Star with the Evening Star, genes with DNA molecules. These, however, turn out not to be unproblematic at all. One may even wonder if any identity claim is ever unproblematic. As the disanalogies, distinctions and perplexities about identity itself began to multiply, the problem of identity took on a life of its own as a logical and metaphysical issue, and the researches no longer illuminated in any specific way the problems of mind. This was partly due to

the diminishing reliance by physicalistically inclined philosophers of mind on any notion of identity at all. The disappearance form of the identity theory is after all misnamed; it is really not an identity theory but only a physicalistic alternative to the identity theory, and as Putman observed in his earliest exposition of functionalism [78], the question of whether to *identify* a logical state of a (realized) Turing machine with its concrete realization in hardware is a relatively idle metaphysical concern. When the two tactics, the disappearance view and functionalism, are put together – as there are strong reasons to do [29, 30] – whatever identities are still left to acknowledge concern rather curious and abstract entities. (For instance, one might be left claiming that the *state of affairs* of Tom's having a pain was identical with the *state of affairs* of his brain being in some particular functional situation.) At this point, the initial motivation for proclaiming *identities*, to save us from the ghostly items of dualism, has vanished.

So it was something of an anachronism when, in 1971, Saul Kripke included a startling “refutation” of the identity theory of mind as an illustrative by-product of his extraordinarily influential, even revolutionary, new account of necessity, designation, and identity [50, 51]. Kripke's argument depends on some technical innovations. A “rigid designator” is an expression that designates the same entity “in all possible worlds.” Thus “Benjamin Franklin” is a rigid designator, while “the inventor of bifocals” is not, though in *this* world they designate the same individual. Kripke argues convincingly that all proper identity claims are composed of terms that are rigid designators and hence when they are true, they are true not contingently, but necessarily. The application to philosophy of mind comes when he argues that expressions such as “my pain” and “my brain state” are both rigid designators, but identity claims composed of them could not be *necessarily* true, hence could not be true at all. (Cf. [62]). Kripke's argument is subtle and ingenious and repays careful study, but it has not commanded assent. Even if one accepts Kripke's new theory of identity and necessity, his arguments to show that the relevant terms are rigid designators are not only vulnerable to straightforward exception, but seem to require assumptions that modern materialists (such as Armstrong) had already been at pains to deny [33, 58]. In retrospect, Kripke's argument can be seen not to have revitalized dualism, but only to have given materialists more and possibly better (deeper) reasons for shunning certain tempting identity claims they had already for the most part learned to avoid.

IV. WHY FUNCTIONALISM?

As befits a view that is the prevailing favorite, functionalism has much to be said for it. Not only does it seem satisfactorily to evade the philosophical objections to all the other forms of materialism, but it is particularly well-suited to serve as the conceptual underpinning for current work in psychology, linguistics, and cybernetics or artificial intelligence. All of these disciplines operate somewhat self-consciously at a certain level of abstraction, and functionalism provides the rationale and justification for just such a strategy. In a way, there is nothing new about it. Psychologists as diverse as Freud and Skinner have shared the basic functionalist tactic: just as Freud eventually realized that any claims he might make about the physical location, composition, or operation of his functionally distinguished entities, the id, ego, and superego, were premature speculations, so Skinner, while granting that no doubt there were *some* internal, physiological mechanisms subserving reinforcement, has abjured speculation or commitment on that score, and settled, with Freud, for mapping the predicted consequences under a variety of circumstances of functionally characterized interactions.

(Skinner, of course, has been less clear than Freud about the fact that he has been committed to a functionalistic model of internal processes. He thinks his peripheralism evades that "charge." That he should worry on this score at all is an embarrassment to philosophers for had he and the other behaviorists not drunk so deeply at the well of logical positivism and the fashionable "operationalism" and "instrumentalism" of that era, they would not have been motivated to constrain their theorizing within such a paralyzing and misguided notion of rigor. Twenty years after logical positivism and even its obituaries have been all but forgotten by philosophers, its dogmas are alive and healthy in the textbooks of behaviorists. It is with some modesty and trepidation, then, that philosophers of mind currently urge their doctrines on their colleagues in other fields. It appears, by the way, that history is about to repeat itself. Now that philosophers of mind have finally succeeded in banishing their fear of internal, "para-mechanical" theoretical entities, a fear they learned from Ryle [27, 39, 93], Roy Schafer [96] has taken Ryle's strictures to heart, and bids fair to initiate an era of Rylean logical behaviorism in psychoanalytic theory. The by now standard arguments against Ryle's behavior-dispositional analyses (they utterly fail to ramify or generalize; they are either obviously false or "saved" by *ceteris paribus* clauses that render them vacuous) seem at least at first glance to transfer intact as criticisms of Schafer.)

More specifically, functionalism provides the conceptual under-pinnings for current work in cognitive psychology, psycholinguistics and artificial intelligence modeling. In these disciplines one research strategy can be characterized in terms of (one version of) Chomsky's distinction between competence and performance: given a specification of a certain sort of competence, say a discriminative competence, or a linguistic competence, the task is to devise a performance model – often a computer simulation program – that exhibits that competence (usually artificially isolated and hedged in various ways) and if possible has a claim to "psychological reality" as well [27, 39]. That is, getting the cat skinned at all can be a major accomplishment; getting it skinned in the way people seem to get it skinned is even better. This sort of research strategy permits highly abstract constraints and difficulties to be explored (how could *anything* learn a natural language? how could *anything* achieve a general capacity for pattern recognition in an unstereotypic environment?) without worrying about the mechanics and the biochemistry of concrete "realizations" in the head, while at the same time not abandoning the fundamental physicalistic constraint that one's functionally described systems be *somehow* physically realizable [25]. At its most general and abstract, this sort of research merges with research in epistemology and philosophy of mind, and it is precisely at this meeting ground that the most promising and exciting work is being done today (e.g. [39, 45, 64, 84, 92, 95]).

Of course functionalism has its skeptics and crisis [11, 12, 47, 56, 57, 66, 90, 100], and the most unsettling problems wear surprisingly traditional garb. First there are problems about the *qualia* of experience, the way it feels to be conscious, and second, there are problems – alluded to earlier – about how a functionally individuated state or event can have *meaning* or *content* in some presumably full-blooded sense required of mental entities. How could a functional state be a pain, and how could a functional state be a thought about *pi*?

V. QUALIA

There is no satisfactory definition of *qualia*, which seems to have become the pet term in the discussions, but the requisite general sense of what qualia are supposed to be is easily captured by an example which figures centrally in the current discussion. Suppose that when you look at a clear "blue" sky you see what I see (insofar as color goes) when I look at a ripe apple, and vice versa, and so forth through all the colors of the spectrum. Your perceived spectrum

is, let us say, a systematic inversion of mine. But since you learned the use of color words just as I did (your parents pointed up at the sky and said "blue," etc.) our use of color words would be indistinguishable, and since we will both associate the color of glowing iron from the blacksmith's hearth with heat, and the color we perceive ice to be with cold, even our secondary descriptions of color (red is a warm color, blue is cool, etc.) might match. In fact, with regard to perceived colors and other qualia such as pains, tickles, sounds, aromas and tastes, do we have any evidence at all, or reason to believe, that any two people experience similar qualia under similar perceptual circumstances? This, the "inverted spectrum" thought experiment, is not new. It was a popular argument among the verificationists who took the "self-evident" unverifiability-in-principle of the hypothesis to mark the meaninglessness of the initial assumption that there are inner sensations or "raw feels" – to use a term philosophy took from the psychologist, Tolman – of the requisite sort at all. As Wittgenstein said, "An 'inner process' stands in need of outward criteria" [117], and although the exegesis of this remark is controversial, it is easy and common to interpret this as an expression of logical behaviorism, an assertion of the incoherence of any doctrine that admits private sensations or experiences of qualia.

It has often been pointed out that today's functionalism is a spiritual descendent of logical behaviorism. Where the logical behaviorists said that being in pain was a matter of behaving or being disposed to behave in a particular way, the functionalists say that being in pain is a matter of being in a functional state of a certain sort – viz., a state that *inter alia* disposes one to certain behavior under certain conditions. From one vantage point, the only difference between the two doctrines appears to be the functionalist's willingness to abandon peripheralism (by countenancing explicitly internal functional states), and the concomitant willingness to define function not just in terms of dispositions to behave, but also in terms of dispositions to change functional state. If, as Skinner says, the skin is not that important a boundary, if internal state-switching counts as behavior, then functionalism is just logical behaviorism in new clothes. It is not surprising then that in the headlong rush away from the verificationism of the recent past, philosophers should attempt to turn the inverted spectrum argument on its head and show that functionalism commits the sin of verificationism of the recent past, philosophers should attempt to turn the inverted spectrum argument on its head and show that functionalism commits the sin of verificationism by failing to grant sense to something that (clever argument reveals) does make sense: the hypothesis of spectrum inversion [11, 12, 56, 100]. In a similar

vein, Nagel has argued [66] that there are certain undeniably meaningful hypotheses about our inner lives and the inner lives of others, about “what it is like to be” a person or a dog or a bat, of which functionalism can give no account.

The strategy of the debate is transparent. The lovers of qualia attempt to establish that functionalism unavoidably *leaves something out*; the wonderful tastes, tones and colors that make life worth living. The functionalists attempt to show that they have not left anything *real* out, and that the alternative to functionalism can only be some unsupportable variety of epiphenomenalism. The issue is not yet resolved, nor will it be resolved by the straightforward victory of one side or the other in a purely conceptual debate. The burden for functionalism is inseparable from the burden of the variety of cognitivistic theories for which it provides the conceptual underpinnings. If an empirical psychological theory develops that is both strongly confirmed and predictive of the rich variety of phenomena of consciousness, we can inspect it for an answer to the question. If it contains a theoretical role for something like qualia, we shall “countenance” qualia in our ontology, but as theoretical entities, not epiphenomena; if no such role appears to be filled, then the very power of the theory will undermine the intuitions that now make the denial of qualia so counterintuitive [26, 29]. If no functionally conceived theory proves up to handling the undisputed facts about consciousness, then that failure of empirical theory, and not any purely philosophical argument, will show that functionalism does in fact leave something out. The philosophical investigations of the issue are not entirely parasitic, however, on the advance of empirical research; they can illuminate the terrain, revealing blind alleys and pitfalls, without attempting to dictate the solution.

VI. INTERNAL REPRESENTATION AND THE PROBLEM OF MEANING

A similar supportive role can be seen for the philosophical contributions to the other main perplexity facing functionalist theories in psychology: the problem of meaning or content. No problems of philosophy have received more, or more expert, attention in recent years than the problems of meaning, and an overview summarizing the work in that area would have to be book length. (There are several fine anthologies of recent work [21, 22, 36, 46, 106].) Almost all of it is at least indirectly relevant to the problems of mind, and some of it is of central importance.

In the late 19th Century Franz Brentano claimed to have discovered the feature that sundered the mental from the physical: Intentionality. Mental

phenomena, he said, differed from physical phenomena in always being directed upon an object, (the object of thought or desire or perception ...) or related to a content (the content of a hope or thought or belief, ...). This was a special sort of relatedness, for the objects of mental phenomena enjoyed a curious sort of “inexistence”. I can want a sloop without there being a sloop I want; the object of my desire is “Intentionally nonexistent” (which does *not* mean deliberately nonexistent: Intentionality has nothing directly to do with what one intends to do). In the 1950s, Chisholm [17] revived Brentano’s notion of Intentionality and (using the tactic of semantic ascent to great effect) turned it into a feature of *language*: the sentence we typically use to talk about mental events have certain peculiarities of logic. Chisholm attempted to characterize those peculiarities of logic in such a way that his distinction between Intentional and non-Intentional *sentences* mirrored Brentano’s distinction between Intentional and non-Intentional *phenomena*. In the ensuing years, Intentionality, viewed as a logical feature of certain classes of propositions, has been exhaustively and fruitfully studied, though it is fair to say that no very broad unanimity has been achieved about the precise definition, status, or role of Intentional discourse [18, 24, 48, 53, 85, 97, 98, 112]. Interest in Intentionality has survived the disagreements and difficulties surrounding its definition not only because it represents an unresolved perplexity of logical theory, but because intuitively it does mark an important divide in our conceptual scheme, though not quite the divide Brentano supposed. The Intentional idioms of our language are roughly those Russell called the idioms of propositional attitude; these idioms typically take “that”-clauses and hence form complex, but not truth-functional, propositions out of others: e.g., “Tom believes that it is raining” contains the proposition “it is raining”, but the truth value of the whole is independent of the truth value of the enclosed proposition. *Roughly*, again, the Intentional idioms are those idioms in our language that relate people, their parts, their acts, their artifacts to *propositions*. Rougher still, we use Intentional idioms to endow things – any sorts of things at all – with *meaning*: if we want to say what Tom believes, what the sentence means, what the frog’s eye tells the frog’s brain, what is innately and tacitly known by the infant language learner [107], what the ego is trying to keep from the superego, what information is stored on the tape, what Houston is signalling to Mariner IV, we use Intentional discourse. A theory of Intentionality, then, would be a theory that made explicit the conceptual ties between these various cognitivistic, information-theoretic, semantic approaches, and that set down the constraints and assumptions involved in the ascription of Intentionally characterized features to things.

That there are very deep problems here can be brought out by considering cognitive psychology.

What unites cognitive psychologists and distinguishes them best from other theorizers in psychology is their willingness to view the individual not simply as a system of functionally individuated parts or subsystems, but of *Intentionality* individuated parts, parts whose functions are to “say that *p*”, “remember that *q*”, “figure out that *r*” – to encode, store, transmit or transform parcels of information. In fact it is their use of Intentional idioms in their science that best marks them off from the behaviorists, in particular Skinner, who has typically misconceived the difference as one between dualists (“mentalists”) and materialists [24, 25, 28]. Cognitive psychologists have generally thought that their use of information-talk is not only proper and well grounded in the mathematical rigor of information-theory and computer science, but positively a great step forward in the fruitful conceptualization of psychology, and so in the end it will be, I believe, but it is not unproblematic. To say that the function of some system is to carry certain information from *a* to *b* is not just like saying the function of the tube is to carry lubricant to the bearing, or the function of the teletype is to convey symbol strings from place to place. Moving information about is not so easily conceived, or if, for some special sense of information, it *is* thus easily conceived, one has purchased the simplicity at the cost of postponing solution to the central problem of Intentionality, as a little thought experiment will show.

Suppose you find yourself locked in a windowless room, with two walls covered with flashing lights, two walls covered with little buttons, and a note telling you that you are imprisoned in the control center of a giant robot on whose safety your own life now depends. Your task is simply to guide the robot through its somewhat perilous environment, learning to discriminate and cope with whatever comes along, finding “nourishment” and safe haven for the robot at night (so you can sleep) and avoiding dangers. All the *information* you need is conveyed by the flashing lights, and the robot’s motor activity is controllable by pushing the buttons. To your dismay, however, you see that none of the lights or buttons are labeled. You can’t tell whether the insistently flashing light in the upper left corner is warning danger, signaling a “full belly,” informing you of the location of the sun, or requesting grease for a heel bearing. You don’t know whether when you push a button and the light goes out, you’ve scratched an itch, occluded your view of something, or destroyed an attacker.

Clearly, if that is all you are given to go on, your task is impossible; if you

succeeded in guiding your robot through the day it would be sheer luck. Yet in one sense (and a very familiar sense to cognitive psychologists) all the information you need is conveyed to you. For we needn't suppose the lights are mere repeaters of peripheral stimulation; their flashing can represent the products of perceptual analysis machinery as sophisticated as you wish, and similarly the output can be supposed to initiate devious actions guided by hierarchical sub-routine systems informed by multi-layered feedback. In short, the entire array of systems devised by the cognitivist psychologists could be built into this robot, so that it conveyed to its control center highly mediated and refined information, and yet, though in one sense the information would be there, in another more important sense, it would not. Yet the task described is in a sense just the brain's task; it has no windows out of which it can look in order to correlate features of the world with its input.

The problem of the control room could be solved for you, of course, if all the lights and buttons were correctly labeled (in a language you knew), but this can hardly be the brain's solution. The job of getting the input information *interpreted* correctly is thus *not* a matter of getting the information translated or transcribed into a particular internal code unless getting the information into that code is *ipso facto* getting it into functional position to govern the behavioral repertoire of the whole organism. This is the problem of meaning that must eventually be faced by any theorist who wishes to appeal to "internal representations" as explicative of psychological phenomena. Some recent work in philosophy directly addresses this issue [23, 27, 29, 39, 45, 64, 84, 92, 95] and this literature depends to various degrees on the fundamental work on meaning that has developed in response to such central themes as Quine's thesis of the indeterminacy of radical translation [85], Austin's work on speech acts [9], and Grice's account of non-natural meaning [40, 41]. A particularly active controversy within the area of internal representation concerns the nature of the supposed vehicles of representation: are they propositional (like sentences) or imagistic or analogical (like pictures or maps) or are there other sorts of "data structures" with no familiar analogues among external vehicles of representation? Work by philosophers in this area merges quite smoothly with that of psychologists and cyberneticists (e.g. [64, 84]), and one can expect this unforced interdisciplinary exchange to produce some genuine advances in outlook in the next few years.

VII. OTHER AREAS OF CURRENT ACTIVITY

This survey of current work is intended to capture the main lines of inquiry, but has left some work unmentioned that is not at all peripheral. For instance, the perplexing status of "introspection" has been carefully studied in a growing literature on "privileged access" and the presumed "incorrugibility" of introspective reports [2, 3, 6, 7, 23, 29, 42, 49, 61, 72, 89, 91, 102, 111], and philosophers have usefully turned their attention to specific uses arising in other fields, such as Sperry's split-brain cases and the various claims being advanced about the different roles of the cerebral hemispheres (e.g. [67]), and the physiology of pain [30, 74, 75]. There has also been a rediscovery of Freud [118], and in particular the problem of self-deception has provoked some excellent work [35, 44, 87, 105]. The "minds and machines" literature has evolved from its early preoccupation with the question "can computers think?" (e.g. [4]) into a much more detailed and informed examination of conceptual issues at the heart of current research in artificial intelligence and automata theory (e.g. [13, 14, 68, 69, 70, 71]). An optimistic prognosis would be that these various strands of inquiry will coalesce into a fairly stable and broadly accepted understanding of the conceptual underpinnings of the functionalistic and physicalistic approach to the mind, but if one attends to the great diversity of opinion in the field, the deep difficulties that can already be seen to attend such a view, and the lesson of history, a more realistic prediction would be that this still fragmentary and tenuous consensus will prove as evanescent as its predecessors and be replaced by a currently unimagined set of doctrines and problems.

NOTE

¹ This overview was originally commissioned by the American Journal of Psychiatry, which eventually declined it on the grounds of being "much beyond" its readership, in spite of my efforts at explanation. I trust that the somewhat elementary tone of the paper is explained by this fact about its originally intended audience. I am indebted to Ned Block, Jeff Titon, Bo Dahlbom, and Sue Stafford for valuable criticisms of earlier drafts of this paper.

REFERENCES

1. Albrton, Rogers: 1968, 'On Wittgenstein's Use of the Term "Criterion"', *Journal of Philosophy*, 56, 845-857. Reprinted in Pitcher, George (ed.): 1968, *Wittgenstein, The Philosophical Investigations*, New York.
2. Alston, William P.: 1971, 'Varieties of Privileged Access', *American*

- Philosophical Quarterly*, 8, 223–251.
3. Alston, William P.: 1976, 'Self-Warrant: A Neglected Form of Privileged Access', *American Philosophical Quarterly*, 13, 257–272.
 4. Anderson, Alan R. (ed.): 1964, *Minds and Machines*, Engelwood Cliffs, NJ.
 5. Anscombe, G.E.M.: 1957, *Intention*, Oxford.
 6. Arbib, Michael A.: 1972, 'Consciousness: The Secondary Role of Language', *Journal of Philosophy*, 69, 579–591.
 7. Armstrong, David M.: 1968, *A Materialist Theory of the Mind*, London.
 8. Austin, John L.: 1961, *Philosophical Papers*, Urmson, J.O. and Warnock, G.J. (eds.), Oxford.
 9. Austin, John L.: 1962, *How to do Things with Words*, Oxford.
 10. Austin, John L.: 1962, *Sense and Sensibilia*, Oxford.
 11. Block, Ned J. and Fodor, Jerry: 1971, 'What Psychological States Are Not', *The Philosophical Review*, 81, 159–181.
 12. Block, Ned J.: 1978, 'Troubles with Functionalism', *Minnesota Studies in the Philosophy of Science*, Vol. 9.
 13. Boden, Margaret: 1974, 'Freudian Mechanisms of Defense: A Programming Perspective', in Wollheim, Richard (ed.): 1974, *Freud: A Collection of Critical Essays*, Garden City, NJ.
 14. Boden, Margaret: 1977, *Artificial Intelligence and Natural Man*, Hassocks, U.K.
 15. Borst, Clive V. (ed.): 1970, *The Mind/Brian Identity Theory*, New York.
 16. Chihara, Charles S. and Fodor, Jerry: 1965, 'Operationalism and Ordinary Language: A Critique of Wittgenstein', *American Philosophical Quarterly*, 2, 281–295. Reprinted in Pitcher, George (ed.): 1968, *Wittgenstein, The Philosophical Investigations*, New York.
 17. Chisholm, Roderick: 1957, *Perceiving: A Philosophical Study*, Ithaca, NY.
 18. Chisholm, Roderick: 1967, 'On some Psychological Concepts and the "Logic" of Intentionality', in Castañeda, Hector-Neri (ed.): *Intentionality, Minds and Perception*, Detroit.
 19. Cornman, James W.: 1971, *Materialism and Sensations*, New Haven, CT.
 20. Davidson, Donald: 1970, 'Mental Events' in Foster, Lawrence and Swanson, Joe W. (eds.): *Experience and Theory*, Amherst, MA.
 21. Davidson, Donald and Harman, Gilbert (eds.): 1972, *The Semantics of Natural Language*, Dordrecht.
 22. Davidson, Donald and Harman, Gilbert: 1975, *The Logic of Grammar*, Encino, CA.
 23. Dennett, Daniel C.: 1969, *Content and Consciousness*, London.
 24. Dennett, Daniel C.: 1971, 'Intentional Systems', *The Journal of Philosophy*, 68, 87–106.
 25. Dennett, Daniel C.: 1975, 'Why the Law of Effect Will Not Go Away', *Journal of the Theory of Social Behaviour*, 5, 169–187.
 26. Dennett, Daniel C.: 1976, 'Are Dreams Experiences?', *The Philosophical Review*, 85, 151–171.
 27. Dennett, Daniel C.: 1975, 'Critical Notice of Jerry Fodor, *The Language of Thought* (New York, 1975)', *Mind*, 86, 265–280.
 28. Dennett, Daniel C.: 1978, 'Skinner Skinned', in Dennett, Daniel C.:

- Brainstorms: Philosophical Essays on Mind and Psychology*, Montgomery, Vermont.
29. Dennett, Daniel C.: 1978, 'Towards a Cognitive Theory of Consciousness', in Savage, C. Wade (ed.): *Minnesota Studies in the Philosophy of Science*, Vol. 9.
 30. Dennett, Daniel C.: 1978, 'Why You Can't Make a Computer that Feels Pain', *Synthese*, **38**, 415–56.
 31. Feigl, Herbert: 1958, 'The "Mental" and the "Physical"', in Feigl, Herbert, Scriven, Michael, and Maxwell, Grover (eds.): *Concepts, Theories and the Mind-Body Problem*, *Minnesota Studies in the Philosophy of Science*, Vol. 2, Minneapolis.
 32. Feigl, Herbert: 1967, *The 'Mental' and the 'Physical' – The Essay and a Postscript*, Minneapolis.
 33. Feldman, Fred: 1974, 'Kripke and the Identity Theory', *The Journal of Philosophy*, **71**, 665–676.
 34. Feyerabend, Paul: 1963, 'Materialism and the Mind-Body Problem' *Review of Metaphysics*, **17**, 49–66.
 35. Fingarette, Herbert: 1969, *Self Deception*, London.
 36. Fodor, Jerry and Katz, Jerrold (eds.): *The Structure of Language*, Engelwood Cliffs, NJ.
 37. Fodor, Jerry: 1965, 'Explanation in Psychology' in Black, Max (ed.) *Philosophy in America*, Ithaca, NY.
 38. Fodor, Jerry: 1968, *Psychological Explanation*, New York.
 39. Fodor, Jerry: 1975, *The Language of Thought*, New York.
 40. Grice, H.P.: 1957, 'Meaning', *The Philosophical Review*, **66**, 377–388.
 41. Grice, H.P.: 1969, 'Utterer's Meaning and Intentions', *The Philosophical Review*, **78**, 147–177.
 42. Gunderson, Keith: 1972, 'Content and Consciousness and the Mind-Body Problem', *The Journal of Philosophy*, **69**, 591–604.
 43. Gustafson, Donald: 1967, *Essays in Philosophical Psychology*, London.
 44. Hamlyn, D.W. and Mounce, H.O.: 1971, 'Self-Deception', *Aristotelian Soc. Supplement*, 45–72.
 45. Harman, Gilbert: 1973, *Thought*, Princeton, NJ.
 46. Harman, Gilbert (ed.): 1974, *On Noam Chomsky: Critical Essays*, Garden City, NJ.
 47. Kölke, William: 1969, 'What is Wrong with Fodor and Putnam's Functionalism', *Nous*, **3**, 83–93.
 48. Kaplan, David: 1968, 'Quantifying in', *Synthese*, **19**, 178–214.
 49. Klein, Barbara V.E.: 1975, 'Some Consequences of Knowing Everything (Essential) There is to Know About One's Mental States', *Review of Metaphysics*, **29**, 3–18.
 50. Kripke, Saul: 1971, 'Identity and Necessity', in Munitz, Milton (ed.), *Identity and Necessity*, New York.
 51. Kripke, Saul: 1972, 'Naming and Necessity', in Davidson, Donald and Harman, Gilbert (eds.), *Semantics of Natural Language*, Dordrecht.
 52. Lewis, David: 1972, 'Psychophysical and Theoretical Identifications', *Australasian Journal of Philosophy*, **50**, 249–258.
 53. Lycan, William, G.: 'On Intentionality and the Psychological', *American*

- Philosophical Quarterly*, **6**, 305–312.
54. Lycan, William G.: 1971, 'Noninductive Evidence: Recent Work on Wittgenstein's Criteria', *American Philosophical Quarterly*, **8**, 109–125.
 55. Lycan, William G.: 1972, 'Materialism and Leibniz' Law', *Monist*, **56**, 276–287.
 56. Lycan, William G.: 1973, 'Inverted Spectrum', *Ratio*, **15**, 315–319.
 57. Lycan, William G.: 1974, 'Mental States and Putnam's Functional Hypothesis', *Australasian Journal of Philosophy*, **52**, 48–62.
 58. Lycan, William G.: 1974, 'Kripke and Materialism', *The Journal of Philosophy*, **71**, 677–689.
 59. Malcolm, Norman: 1956, 'Dreaming and Skepticism', *The Philosophical Review*, **65**, 14–37.
 60. Malcolm, Norman: 1959, *Dreaming*, London.
 61. Margolis, Joseph: 1970, 'Indubitably, Self-Intimating Sates and Logically Privileged Access', *The Journal of Philosophy*, **67**, 918–931.
 62. Matson, Wallace: 1976, *Sentience*, Berkeley, CA.
 63. Melden, A.I.: 1961, *Free Action*, London.
 64. Minsky, Marvin: 1973, 'Frame Systems: A Framework for Representing Knowledge', MIT Artificial Intelligence Lab Report.
 65. Nagel, Thomas: 1965, 'Physicalism', *The Philosophical Review*, **74**, 339–356.
 66. Nagel, Thomas: 1971, 'Brain Bisection and the Unity of Consciousness', *Synthese*, **22**, 396–413.
 67. Nagel, Thomas: 1974, 'What is it Like to be a Bat?', *The Philosophical Review*, **83**, 435–450.
 68. Nelson, Raymond: 1975, 'Behaviorism, Finite Automata and Stimulus Response Theory', *Theory and Decision*, **6**, 249–267.
 69. Nelson, Raymond: 1975, 'On Machine Expectation', *Synthese*, **31**, 129–139.
 70. Nelson, Raymond: 1976, 'On Mechanical Recognition', *Philosophy of Science*, **43**, 24–52.
 71. Nelson, Raymond: 1976, 'Mechanism, Functionalism and the Identity Theory', *The Journal of Philosophy*, **73**, 365–385.
 72. Parsons, Kathryn P.: 1970, 'Mistaking Sensation', *The Philosophical Review*, **79**, 201–213.
 73. Peters, Richard: 1958, *The Concept of Motivation*, London.
 74. Pitcher, George: 1970, 'Pain Perception', *The Philosophical Review*, **79**, 368–393.
 75. Pitcher, George: 1970, 'The Awfulness of pain', *The Journal of Philosophy*, **67**, 481–492.
 76. Place, U.T.: 1956, 'Is Consciousness a Brain Process?', *British Journal of Psychology*, **42**, 44–50.
 77. Price, H.H.: 1932, *Perception*, London.
 78. Putnam, Hilary: 1960, 'Minds and Machines', in Hook, Sidney (ed.), *Dimensions of Mind*, New York. Reprinted in [83].
 79. Putnam, Hilary: 1962, 'Dreaming and "Depth Grammar"', in Butler, R.J. (ed.), *Analytical Philosophy*, Oxford.
 80. Putnam, Hilary: 1964, 'Robots, Machines or Artificially Created Life', *The Journal of Philosophy*, **61**, 668–691.

81. Putnam, Hilary: 1967, 'Psychological Predicates', in *Art, Mind, and Religion*, Pittsburgh. Reprinted as 'The Nature of Mental States' in Putnam, Hilary: 1975, *Mind, Language and Reality* (Philosophical Papers, Vol. II), Cambridge.
82. Putnam, Hilary: 1967, 'The Mental Life of Some Machines' in Castañeda, Hector-Neri (ed.): *Intentionality, Minds and Perception*, Detroit. Reprinted in [83].
83. Putnam, Hilary: 1975, *Mind, Language and Reality* (Philosophical Papers, Vol. II), Cambridge, MA.
84. Pylyshyn, Zenon: 1964, 'What the Mind's Eye Tells the Mind's Brain', *Psychol. Bull.*, **80**, 1–24.
85. Quine, W.V.O.: 1960, *Word and Object*, Cambridge, MA.
86. Quine, W.V.O.: 1969, 'Epistemology Naturalized', in *Ontological Relativity and Other Essays*, New York.
87. Rorty, Amelie O.: 1972, 'Belief and Self-Deception', *Icarus*, **15**, 387–410.
88. Rorty, Richard: 1965, 'Mind-Body Identity, Privacy and Categories', *Review of Metaphysics*, **19**, 24–54.
89. Rorty, Richard: 1970, 'Incorrígibility as the Mark of the Mental', *Journal of Philosophy*, **67**, 399–424.
90. Rorty, Richard: 1972, 'Functionalism, Machines and Incorrígibility', *Journal of Philosophy*, **69**, 203–220.
91. Rorty, Richard: 1972, 'Dennett on Awareness', *Philosophical Studies*, **23**, 153–162.
92. Rosenberg, Jay: 1974, *Linguistic Representation*, Dordrecht.
93. Ryle, Gilbert: 1949, *The Concept of Mind*, London.
94. Ryle, Gilbert: 1958, 'A Puzzling Element in the Notion of Thinking', *Proceedings of the British Academy* **44**, 129–144. Reprinted in Strawson, Peter F. (ed.): 1968, *Studies in the Philosophy of Thought and Action*, Oxford.
95. Savage, C. Wade (ed.): *Minnesota Studies in the Philosophy of Science*, Vol. 9. 1978.
96. Schafer, Roy A.: 1975, *A New Language for Psychoanalysis*, New Haven, CT.
97. Sellars, Wilfrid: 1963, *Science, Perception and Reality*, London.
98. Sellars, Wilfrid: 1964, 'Notes on Intentionality', *The Journal of Philosophy*, **61**, 655–666.
99. Shaffer, Jerome: 1961, 'Could Mental States Be Brain Processes?', *The Journal of Philosophy*, **58**, 813–822.
100. Shoemaker, Sidney: 1975, 'Functionalism and Qualia', *Philosophical Studies*, **27**, 291–315.
101. Smart, J.J.C.: 1959, 'Sensations and Brain Processes', *The Philosophical Review*, **68**, 141–156.
102. Smart, J.J.C.: 1962, 'Brain Processes and Incorrígibility', *Australasian Journal of Philosophy*, **40**, 68–70.
103. Smart, J.J.C.: 1963, *Philosophy and Scientific Realism*, London.
104. Smart, J.J.C.: 1970, 'Critical Notice of Content and Consciousness by D.C. Dennett', *Mind*, **79**, 616–623.
105. De Sousa, Ronald: 1970, 'Self-Deception', *Inquiry*, **13**, 308–334.
106. Steinberg, Danny D. and Jakobovitz, Leon A. (eds.): 1971, *Semantics*, Cambridge.

107. Stich, Steven (ed.): 1975, *Innate Ideas*, Berkeley, CA.
108. Strawson, Peter F.: 1959, *Individuals*, London.
109. Taylor, Brandon: 1973, 'Mental Events: Are There Any?', *Australasian Journal of Philosophy*, **51**, 189–200.
110. Taylor, Charles: 1967, 'Mind-Body Identity, A Side Issue?', *The Philosophical Review*, **76**, 201–213.
111. Tormey, Alan: 1973, 'Access, Incorrugibility and Identity', *The Journal of Philosophy*, **70**, 115–128.
112. Troyer, John G. and Wheeler, Samuel (eds.): 1974, 'Intentionality, Language and Translation', *Synthese*, **23** 123–456.
113. Urmson, J.O.: 1952, 'Motives and Causes', *Aristotelian Soc. Suppl.* **26** 179–194.
114. Warnock, G.J.: 1954, 'Seeing', *Aristotelian Soc. Proc.*, **55** 201–218.
115. White, Alan R.: 1959–1960, 'Different Kinds of Need Concepts', *Analysis*, **20**, 112–116.
116. Wisdom, John: 1946, 'Other Minds', *Aristotelian Soc. Suppl.*, **20**, 122–147.
117. Wittgenstein, Ludwig: 1953, *Philosophical Investigations*, Oxford.
118. Wollheim, Richard (ed.): 1974, *Freud: A Collection of Critical Essays*, Garden City, NJ.

Tufts University

Received September 19, 1977

PART I

COMPUTATIONAL CONCEPTIONS

MACHINES AND THE MENTAL*¹

Computers are machines and there are a lot of things machines can't do. But there are a lot of things *I* can't do: speak Turkish, understand James Joyce, or recognize a nasturtium when I see one. Yet, numerous as are my disabilities, they do not materially affect my status as a thinking being. I lack specialized skills, knowledge and understanding, but nothing that is essential to membership in the society of rational agents. With machines, though, and this includes the most sophisticated modern computers, it is different. They *do* lack something that is essential.

Or so some say. And so say I. In saying it, though, one should, as a philosopher, be prepared to say what *is* essential, what are the conditions for membership in this exclusive club. If an ability to understand James Joyce isn't required, what, then, *must* one be able to understand? If one doesn't have to know what nasturtiums look like, is there something else one must be able to identify? What might this be? If one is told that there is no specific thing one has to understand, identify or know but, nonetheless, something *or other* towards which one must have a degree of competence, it is hard to see how to deny computers admission to the club. For even the simple robots designed for home amusement talk, see, remember and learn. Or so I keep reading in the promotional catalogs. Isn't this enough? Why not?

I happen to be one of those philosophers who, though happy to admit that minds compute, and in this sense *are* computers, have great difficulty seeing how computers could be minded. I'm not (not *now* at least) going to complain about the impoverished inner life of the computer – how they don't feel pain, fear, love or anger. Nor am I going to talk about the mysterious inner light of consciousness. For I'm not at all sure one needs feelings or self consciousness to solve problems, play games, recognize patterns and understand stories. Why can't pure thought, the sort of thing computers purportedly have, stand to ordinary thought, the sort of thing we have, the way a solitary stroll stands to a hectic walk down a crowded street? The same thing – walking – is going on in both cases. It just *seems* different because, in the latter case, so much else is going on at the same time. A mathematician's calculations are no less brilliant, certainly no less deserving of classification as mental, because he or she is blind, deaf and emotionally stunted – because, in other words, the calculations occur within a comparatively anemic sensory

and emotional environment. Why can't we think of our machines as occupying a position on the far right of this mental continuum? Just a bit to the right of Star Trek's Doctor Spock? We don't, after all, deny someone the capacity for love because they can't do differential calculus. Why deny the computer the ability to solve problems or understand stories because it doesn't feel love, experience nausea, or suffer indigestion?

Nor am I going to talk about how bad computers are at doing what most children can do – e.g., speak and understand their native language, make up a story or appreciate a joke. For such comparisons make it sound like a competition, a competition in which humans, with their enormous head start, and barring dramatic breakthroughs in AI, will remain unchallenged for the foreseeable future. I don't think the comparison should be put in these terms because I don't think there is a genuine competition in this area at all. It isn't that the best machines are still at the level of two-year-olds, requiring only greater storage capacity and fancier programming to grow up. Nor should we think of them as idiot savants, exhibiting a spectacular ability in a few isolated areas, but having an overall IQ too low for fraternal association. For machines, even the best of them, don't have an IQ. They don't *do* what we do – at least none of the things that, when we do them, exhibit intelligence. And it's not just that they don't do them the way we do them or as well as we do them. They don't do them at all. They don't solve problems, play games, prove theorems, recognize patterns, let alone think, see and remember. They don't even add and subtract.

To convince you of this, it is useful to look at our relationship to various instruments and tools. The preliminary examination will not take us far, but it will set the stage for a clearer statement of what I take to be the fundamental difference between minds and machines.

In our descriptions of instruments and tools we tend to assign them the capacities and powers of the agents who use them. We often think, or at least talk, of artifacts – tools, instruments and machines – as telling us things, recognizing, sensing, remembering and, in general, doing things that, in our more serious, literal, moments, we acknowledge to be the province of rational agents. In most cases this figurative use of language does no harm. No one is really confused. Though we open doors, and keys open (locked) doors, no one seems to worry about whether keys open doors better than we do, whether we are still ahead in this competition. No one is trying to build a fifth generation key that will surpass us in this enterprise. Why not? Since both keys and people open doors, why doesn't it make sense to ask who does it better? Because, of course, we all understand that doors are opened *with* keys.

We are the agents. The key is the instrument. That we sometimes speak of the instrument in terms appropriate to the agent, speak of the key as doing what the agent does with the key, should not tempt us into supposing that, therefore, there are some things we do that keys can also do. We catch fish with worms; *we*, not the worms, catch the fish.

Before concluding, however, that computers are, like keys, merely fancy instruments in our cognitive tool box – and, thus, taken by themselves, unable to do what we can do with them – consider another case. Who really picks up the dust, the maid or the vacuum cleaner? Is the vacuum cleaner merely an instrument that the maid uses to pick up dust? Well yes, but not *quite* the way one uses a key to open a door or a hammer to pound a nail. One pushes the vacuum cleaner around but *it* picks up the dust. In this case (unlike the key case) the question: “Who picks up dust better: people or vacuum cleaners?” *does* make good sense, and the answer, obviously, is the vacuum cleaner. We may never have had any real competition from keys for opening doors, but we seem to have lost the race for picking up dust to vacuum cleaners.

What such examples reveal is that the agent–instrument distinction is no certain guide to what is to be given credit for a performance. We do things. Machines do things. Sometimes we do things with machines. Who gets the credit depends on what is done and how it is done. To ask whether a simple pocket calculator can really multiply or whether it is *we* who multiply *with* the calculator is to ask, whether, relative to this task, the agent–instrument relation is more like our use of a key in opening a door more like our use of a vacuum cleaner in picking up dust.

Well, then, are computers our computational keys? Or are they more like vacuum cleaners? Do they literally do the computational tasks that we sometimes do without them but do it better, faster, and more reliably? This may sound like a rather simple-minded way to approach the issue of minds and machines, but unless one gets clear about the relatively simple question of *who* does the job, the person or the pocket calculator, in adding up a column of figures, one is unlikely to make much progress in penetrating the more baffling question of whether more sophisticated machines exhibit (or will some day) some of the genuine qualities of intelligence. For I assume that if machines can really play chess, prove theorems, understand a text, diagnose an illness, and recognize an object – all achievements that are routinely credited to modern machines by sober members of the artificial intelligence community – if these descriptions are *literally* true, then to that degree they participate in the intellectual enterprise. To that degree they are minded. To that extent they belong in the club however much *we*, with our

prejudice in favor of biological look-alikes, may continue to deny them full admission.

So let me begin with a naive question: Can computers add? We may not feel very threatened if this is *all* they can do. Nevertheless, if they do even this much, then the barriers separating mind and machine have been breached and there is no reason to think they won't eventually be removed.

The following argument is an attempt to show that whatever it is that computers are doing when we use them to answer our arithmetical questions, it isn't addition. Addition is an operation on numbers. We add 7 and 5 to get 12, and 7, 5 and 12 are numbers. The operations computers perform, however, are not operations on numbers. At best, they are operations on certain physical tokens that *stand for*, or are interpreted as standing for, the numbers. Therefore, computers don't add.

In thinking about this argument (longer than I care to admit) I decided that there was something right about it. *And* something wrong. What is right about it is the perfectly valid (and relevant) distinction it invokes between a representation and what it represents, between a sign and what it signifies, between a symbol and its meaning or reference. We have various ways of representing or designating the numbers. The written numeral "2" stands for the number 2. So does "two." Unless equipped with special pattern recognition capabilities, machines are not prepared to handle these particular symbols (the symbols appear on the keyboard for *our* convenience). But they have their own system of representation: open and closed switches, the orientation of magnetic fields, the distribution of holes on a card. But whatever the form of representation, the machine is obviously restricted to operations on the symbols or representations themselves. It has no access, so to speak, to the *meaning* of these symbols, to the things the representations represent, to the numbers. When instructed to add two numbers stored in memory, the machine manipulates representations in some electromechanical way until it arrives at another representation-something that (if things go right) stands for the sum of what the first two representations stood for. At no point in the proceedings do numbers, in contrast to numerals, get involved. And if, in order to add two numbers, one has to perform some operation on the numbers themselves, then what the computer is doing is not addition at all.

This argument, as I am sure everyone is aware, shows *too* much. It shows that *we* don't add either. For whatever operations may be performed in or by our central nervous system when we add two numbers, it quite clearly isn't an operation on the numbers themselves. Brains have their own coding

systems, their own way of representing the objects (including the numbers) about which its (or our) thoughts and calculations are directed. In this respect a person is no different than a computer. Biological systems may have different ways of representing the objects of thought, but they, like the computer, are necessarily limited to manipulating these representations. This is merely to acknowledge the nature of thought itself. It is a *vicarious* business, a *symbolic* activity. Adding two numbers is a way of thinking *about* two numbers, and thinking *about* *X* and *Y* is not a way of pushing *X* and *Y* around. It is a way of pushing around their symbolic representatives.

What is wrong with the argument, then, is the assumption that in order to add two numbers, a system must literally perform some operation on the numbers themselves. What the argument shows, if it shows anything, is that in order to carry out arithmetical operations, a system must have a way of representing the numbers and the capacity for manipulating these representations in accordance with arithmetic principles. But isn't this precisely what computers have?

I have discussed this argument at some length only to make the point that all cognitive operations (whether by artifacts or natural biological systems) will necessarily be realized in some electrical, chemical or mechanical operation over physical structures. (Or, if materialism isn't true, they will be realized in or by transformations of mind-stuff.) This fact alone doesn't tell us anything about the cognitive nature of the operations being performed – whether, for instance, it is an inference, a thought or the taking of a square root. For what makes these operations into thoughts, inferences, or arithmetical calculations is, among other things, the meaning or, if you prefer, the semantics of those structures over which they are performed. To think about the number 7 or your cousin George, you needn't do anything with the number 7 or your cousin George, but you do need the internal resources for representing 7 and George and the capacity for manipulating these representations in ways that stand for activities and conditions of the things being represented.

This should be obvious enough. Opening and closing relays doesn't count as addition, or as moves in a chess game, unless the relays, or their various states, stand for numbers and chess moves. But what may not be so obvious is that these physical activities cannot acquire the relevant kind of meaning merely by *assigning* them an interpretation, by letting them mean something *to* or *for us*. Unless the symbols being manipulated mean something *to the system manipulating them*, their meaning, whatever it is, is irrelevant to evaluating what the system is doing when it manipulates them.² I cannot

make you, someone's parrot or a machine think about my cousin George, or the number 7, just by assigning meanings in accordance with which this is what your (the parrot's, the machine's) activities stand for. If things were this easy, I could make a tape recorder think about my cousin George. Everything depends on whether this is the meaning these events have to you, the parrot, or the machine.

Despite some people's tendency to think that the manipulation of symbols is *itself* a wondrous feat, worthy of such inflated descriptions as "adding numbers," "drawing conclusions," or "figuring out its next move" the process is, in fact, absolutely devoid of cognitive significance.³ I once watched a gerbil manipulate a symbol, a symbol that, according to conventional standards – standards that I, but not the gerbil – understood, stood for my bank balance. I didn't have the slightest temptation to see in this symbol manipulation (actually *consumption*) process anything of special significance. Even if I trained a fleet of gerbils to arrange symbols in some computationally satisfying way (e.g., to balance my checkbook), I don't think *they* should be credited with balancing my checkbook. *I* would merely be using the gerbils to balance my checkbook in the way I use worms to catch fish.

To understand *what* a system is doing when it manipulates symbols, it is necessary to know, not just what these symbols mean, what interpretation they have been, or can be, *assigned*, but what they mean to the system performing the operations. John Searle and Ned Block have dramatized this point.⁴ Searle, for instance, asks one to imagine someone who understands no Chinese manipulating Chinese symbols in accordance with rules expressed in a language he does understand. Imagine the rules cleverly enough designed so that this person can carry on a correspondence in Chinese – responding to (written) Chinese questions with (written) Chinese answers in a way that is indistinguishable from the performance of a native speaker of Chinese. Clearly, though a correspondent might not be able to discover this fact, the symbol manipulator himself doesn't understand Chinese. Nor does the system of which he is a part. Understanding Chinese is not just a matter of manipulating meaningful symbols in some appropriate way. These symbols must mean something *to* the system performing the operations.

This should not be taken to imply that machines cannot serve as useful models for cognitive processes. On the contrary. Their prevalent use in cognitive psychology indicates otherwise. What it does imply is that the machines do not literally *do* what we do when we engage in those activities for which they provide an effective model. Computer simulations of a hurricane do not blow trees down. Why should anyone suppose that computer

simulations of problem solving must themselves solve problems?

But how does one build a system that is capable not only of performing operations on (or with) symbols, but one *to which* these symbols mean something, a machine that, in this sense, understands the meaning of the symbols it manipulates? Only when we can do this will we have machines that not only produce meaningful output, but machines whose activities in producing that output bear the mark of the mental. Only then will we have machines that we can not only *use* to balance our checkbook, but machines that will do it for us, machines that will not only print out answers to our questions, but machines that will *answer* our questions.

One thing seems reasonably clear: if the meaning of the symbols on which a machine performs its operations is a meaning wholly derived from us, its users – if it is a meaning that *we* assign the various states of the machine and, therefore, a meaning that we can change at will without altering the *way* these symbols are processed by the machine itself – then there is no way the machine can acquire understanding, no way these symbols can have a meaning to *the machine itself*. Unless these symbols have what we might call an intrinsic meaning, a meaning they possess which is independent of our communicative intentions and purposes, then this meaning *must* be irrelevant to assessing what the machine is doing when it manipulates them. The machine is processing meaningful (to us) symbols, to be sure, but the *way* it processes them is quite independent of *what they mean* – hence, nothing *the machine* does is explicable in terms of the meaning of the symbols it manipulates or, indeed, of their even having a meaning. Given the right programming and data base, we can contrive to make the sentences a machine produces answers to our questions. But the machine itself is no more answering our questions than is an automatic teller (now so prevalent in the banking industry) embezzling money when it keeps our deposit without crediting our account.

In order, therefore, to approximate something of genuine cognitive significance, in order to give a machine something that bears a mark, if not *all* the marks, of the mental, the symbols a machine manipulates must be given a meaning of their own, a meaning that is independent of their user's purposes and intentions. Only by doing this will it become possible to make the meaning of these symbols relevant to what the machine does with them, possible, in other words, to make the machine do something *because* of what its symbols mean, possible, therefore, to make these symbols mean something to the machine itself.

And how might this be done? In the same way, I submit, that nature

arranged it in our case. We must put the computer into the head of the robot, into a larger system that has the kind of sensory capabilities, the perceptual resources, that enable what goes on inside the computer to mean something, in Paul Grice's natural sense of meaning,⁵ about what goes on outside the computer. The elements over which the computer performs its operations will then have a meaning that is independent of the conventions of its users. They will then mean something in the same way the swing of a galvanometer needle means something regarding the electrical activity in the circuit to which it is connected, the way expanding mercury means something about the surrounding temperature, the way a voltage spike in our visual cortex means something about the distribution of light impinging on the retina. This kind of meaning is sometimes called information.⁶ It is the kind of meaning we associate with reliable signs and trustworthy indicators, the kind of meaning possessed by dark clouds, shadows, prints, leaf patterns, smoke, acoustic vibrations, and the electrical activity in the sensory pathways. The difference between a robot and the disembodied computer found in our office buildings and laboratories is that the former, unlike the latter, have symbol systems that are also *sign* systems: signs being symbols having a meaning quite independent of what *we* might say or think they mean. The only intrinsic meaning in most computers is the meaning derived from the array of pressure sensitive transducers on its keyboard. The activities in the computer may mean a move to KB-3 *to us*, but all they mean *to the computer* is that key 37 has been depressed.

This is only to say that information, *real* information, the kind of meaning associated with natural signs, is irrelevant to the operation of high speed digital computers in a way it is not irrelevant to the operation of living systems. If a sea snail doesn't get information about the turbulence in the water, if there isn't some state *in* the snail that functions as a natural sign of turbulent water, it risks being dashed to pieces when it swims to the surface to obtain the micro-organisms on which it feeds. If (certain) bacteria did not have something inside that meant that *that* was the direction of magnetic north, they could not orient themselves so as to avoid toxic surface water. They would perish. If, in other words, an animal's internal sensory states were not rich in information, intrinsic natural meaning, about the presence of prey, predators, cliffs, obstacles, water and heat it could not survive. It isn't enough to have the internal states of these creatures mean something *to us*, for it to have *symbols* it can manipulate. If these symbols don't somehow register the conditions in their possessor's surroundings, the creature's symbol manipulation capacity is completely worthless. Of what possible

significance is it to be able to handle symbols for food, danger and sexual mates if the occurrence of these symbols is wholly unrelated to the actual presence of food, danger and sexual mates?

In a sense, then, work on machine perception, pattern recognition, and robotics has greater relevance to the cognitive capacities of machines than the most sophisticated programming in such purely intellectual tasks as language translation, theorem proving, or game playing. For a pattern recognition device is at least a device whose internal states, like those of the bacterium, snail and human being, mean something about what is happening, or the conditions that exist, around it. There is actually something *in* these machines that means something regarding what is happening outside them and, moreover, something that means this whether or not we, the users of the machine (or, indeed, the machine itself), recognize it. We are not free to assign or withhold this meaning – anymore than we are free to say what the screech of a smoke alarm means. We can *say* that the alarm means there are leopards nearby, and for certain purposes (e.g., in a children's game of make-believe) we may even want to give it that meaning. But that isn't actually what the sound means. That isn't what it is a sign of, not the information it carries. And for the same reason, the meaning of the internal states of a pattern recognition device, or a robot equipped with sensory capacities, is a meaning these states have which, if it isn't actually a meaning *for* the machine itself, is the only meaning that shows any promise of being promoted into something that is relevant to assessing what these machines are doing when they mobilize these meaningful elements to produce an output.

But have we come any closer to understanding genuine mentation, the capacity to add, subtract, play games, understand stories, and think about one's cousin George? What we have so far required of any aspiring symbol-manipulator is, in effect, that *some* of its symbols be actual signs of the conditions they signify, that there *be* some system-to-world correlations that confer on these symbols an intrinsic meaning, a meaning they do not derive wholly from the purposes and intentions of their users. This puts the symbol-manipulator *in the world* in a way it would not otherwise be. But have we come any closer to understanding how an element, symbol *or* sign, could have meaning *to* the symbol manipulator itself, how this meaning, and not just the sign having this meaning, could be relevant to *what* the system is doing when it manipulates these signs?

Think about a dog that has been trained to detect marijuana. Customs' agents can use these dogs to find concealed marijuana. When the dog barks, wags its tail, or does whatever it was trained to do when it smells marijuana,

this alerts the agent to its presence. As a result of the dog's behavior, the official comes to believe that there is marijuana in the suitcase. But what does *the dog* believe? Surely not what the agent believes — *viz.*, that there is marijuana in the suitcase. Why not? There is obviously something in the dog that is sensitive to the presence of marijuana, some neural condition whose occurrence is a sign, and in this sense means, that there is marijuana nearby. Furthermore, this something is (as a result of training) getting the dog to wag its tail or bark. Why isn't this enough to justify attributing a belief to the dog, a belief with the content: there is marijuana nearby? If we had a Stanford robot that could perform half as well with blocks on a table, we would doubtless be hearing about its extraordinary recognitional capacities. But nobody seems terribly impressed with the dog. The dog, one can hear them saying, has a wonderfully discriminating sense of smell. It has *sensory* powers that exceed those of its trainers, but its *conceptual* or *cognitive* capacities are modest indeed. It can smell marijuana, sure enough. It can even be trained to respond in some distinctive way to this smell. But it doesn't have the conceptual resources for believing *of* what it smells *that* it is marijuana.

If we are going to treat the dog in this deflationary way, we should be prepared to do the same with machines — including fancy robots. In industrial applications of machine vision, for example, it is said that machines can recognize short circuits on the printed circuit boards they examine. Not so. The machine merely searches for breaks or discontinuities in the metallic deposit. It is concerned with *spatial* discontinuities. We, its users, are worried about electrical discontinuities. Under the right circumstances, we can use something that detects the first as an instrument, a means, for identifying the second, but, just like the dog, the instrument should not be credited with the *conceptual* talents of its users, what *we* are able to discover by using it. The machine is no more able to have electricity thoughts than the dog is able to have marijuana thoughts.

Some people think that what machines lack is conscious awareness. Perhaps they do. But our marijuana sniffing dog should teach us that this isn't the missing ingredient, not what we need to manufacture a thinker of thoughts out of a sign manipulator. For the dog *is*, whereas the custom's agent is not, *aware* of the concealed marijuana. The dog smells it and the agent does not. To give a system the kind of meaning we now seek, to give it genuine understanding, it is not enough to give it conscious awareness *of* the stuff it is supposed to cognize. It isn't even enough to make the creature's conscious awareness of the stuff *cause* it to behave in some appropriate way

toward the stuff. For, as our trained dog illustrates, all this can be true without the system, animal or machine, having the slightest conception of what it all means. And what we are after, of course, is something that wags its tail, activates its printer, or starts its motors, not just because it is aware of, say, marijuana, but because it thinks, judges, or believes that it is marijuana. What we are after is *conception*, not *perception*.

The difference between machines (or dogs) and the agents who use them is that although machines (and dogs) can pick up, process and transmit the information we need in our investigative efforts (this is what makes them useful tools), although they can respond (either by training or programming) to meaningful signs, it isn't the meaning of the signs that figures in the explanation of why they do what they do. Some internal sign of marijuana, some neurological condition that, in this sense, means that marijuana is present, can cause the dog's tail to move, but it isn't the fact that it means this that explains the tail movement. This, I submit, is the difference between the dog and its master, between the machine and its users, between the robot and the people it replaces. When I smell marijuana, my finger wagging is produced, not simply (as in the case of the dog) by a neurological condition that means that marijuana is present, but by the *meaning* of this neurological condition, by the fact that it means this and not something else. In *my* case the motor activity is produced by the meaning of an occurrent sign; in the dog's case by the occurrence of a sign having that meaning. To say that the smell of marijuana means something to me that it doesn't mean to the dog is merely to say that its meaning what it does makes a difference to what I do but not to what the dog does. That is why it is true of me, but not the dog, that I wag my finger because I think marijuana is present, because I am in an internal state having this content. The dog is in a state with the same content, to be sure, but it isn't this content that wags the tail. The difference between a thinker of marijuana thoughts (me) and the mere detector of marijuana (dog or machine) is not, then, merely a difference in what our internal signs mean, but a difference in whether, and if so, how, these meanings are implicated in the management of the signs themselves.

I seem to have painted myself into a corner. At least I expect to be told as much by those philosophers who are deeply suspicious of meaning. I expect to be told that meaning is an abstraction, not something that *could* play a role in the activities of a symbol manipulating system. From the control point of view, meaning is an epiphenomenon. It is causally inert. Even if one agrees that there are signs in the head, it is the signs themselves, not their meaning, that turn the cranks, pull the levers, and depress the accelerator. It is the grey

stuff inside, not what it means, that activates the motor neurons. Just ask the neurobiologists. If, in order to promote a processor of meaningful signs into a system with genuine understanding, into a real thinker of thoughts, we must give the meaning of these signs a role to play in the *way* these signs are processed, in the way the motor control system operates with them, then the prospects for effecting such a promotion, not just for machines, but for human beings as well, look bleak indeed.

Such pessimism, though widespread these days, is unwarranted. Meanings, of the kind now in question, *are* what philosophers like to call abstract entities, but they are *no more* abstract, and certainly no less capable of exercising a causal influence, than are, say, differences in weight, brightness and orientation. Just as the difference in weight between a basketball and a bowling ball may be responsible, causally responsible, for the behavior of a beam balance, the correlations constituting the meaning of a sign can, and regularly *do*, affect the way a system processes that sign. The correlation between a ringing bell and someone's presence at one's door, the kind of correlation that confers on the ringing bell the meaning that someone is at the door, *changes* the way a (suitably exposed) nervous system processes the internal sign of a ringing bell. Exposure (either directly or indirectly) to this correlation produces a difference in whether, and if so, which, motor neurons are activated by the internal sensory sign of a ringing bell. This, it seems to me, is a case where the meaning of a sign, and not just the sign that has that meaning, makes a difference in *how* a system processes that sign – hence, a case where the sign's meaning, and not the sign itself, helps to explain the behavior of the system in which that sign occurs.

My doorbell example is a homely example of the causal role of meaning. Some may think it ignores all the interesting questions. For it involves an agent already possessed of the conceptual resources for interpreting signs, understanding meanings, and modifying his or her behavior in the light of experienced correlations. This is true, but irrelevant. For the very same phenomenon can be illustrated at almost every biological level, every level at which *learning* occurs. It is, in fact, merely an instance of what learning theorists describe as the contingencies modifying the way a system processes, and hence responds to, the internal signs for stimulus conditions. Even the lowly snail mentioned earlier changes the way it processes signs by exposing it to the correlations constituting the meaning of these signs. And it is surely, the fact that our internal states are correlated with certain kinds of external conditions that helps to determine the ultimate outcome of the motor activities produced by these internal states. It is the correlations, therefore,

that help to determine what kind of feedback we received from such activities and, hence, the likelihood of our repeating them in the same circumstances. It is, therefore, the correlations, not merely the internal correlates, that shape – hence *explain* – learned behavior. Learning, in fact, *is* a process in which the meaning of internal signs (i.e., their correlation with external conditions), not (merely) the signs themselves, helps to determine how these signs are exploited for purposes of motor control. For such systems the internal signs not only have meaning, this meaning affects the way the system manages these signs; and it is in this sense that the signs mean something *to* the system in which they occur.

This, it seems to me, is a fundamental difference between the sign processing capabilities of various systems. It is a difference that helps explain why it seems so natural to say of some of them (human beings and some animals) but not others (machines and simple organisms) that the symbols they manipulate mean something to the symbol manipulator. It is a difference, I submit, that underlies our conviction that we, but not the machines and a variety of simple organisms, are genuine thinkers of thoughts. What gives us the capacities underlying this difference is a long and complicated story. It involves, I think, issues in learning theory, our multiple sensory access to the things we require to satisfy our needs, and the kind of feedback mechanisms we possess that allow us to modify *how* we manipulate internal signs by the kind of results our previous manipulations have produced. But this, clearly, is a story that we expect to hear from neurobiologists, not from philosophers. All I have been trying to tell is a simpler story, a story about the entrance requirements for admission to the club. I leave it to others to worry about how different systems manage, each in their own way, to satisfy these requirements.

NOTES

* Presidential Address delivered before the Eighty-third Annual Meeting of the Western Division of the American Philosophical Association, Chicago, Illinois, April 26, 1985.

¹ My thanks to Denny Stampe for careful criticism and many useful suggestions. I also want to acknowledge the help given me by Fred Adams and the other sceptics in the audience at Augustana College where I read an early draft of this paper. They convinced me that the draft I read them was *earlier* than I ever suspected.

² This is what Haugeland calls “original intentionality,” something that, according to Haugeland, computers don’t have: “To put it bluntly: computers themselves don’t mean anything by their tokens (any more than books do) – they only mean what we

say they do. Genuine understanding, on the other hand, is intentional in its own right" and not derivatively from something else. *Mind Design*, John Haugeland (ed.), Bradford Brooks; Montgomery, Vt., 1981, pp. 32–33. A number of authors have made essentially this point in their own way; e.g., Jerry Fodor, 'Tom Swift and His Procedural Grandmother', *Cognition*, 6, (1978), reprinted in *Representations*, MIT/Badford, 1981; Hilary Putnam, 'Brains in a Vat', *Reason, Truth and History*, Cambridge University Press, 1981, pp. 10–11; Rob Cummins, *The Nature of Psychological Explanation*, MIT/Badford, 1983, p. 94; Tyler Burge, 'Belief *De Re*', *The Journal of Philosophy*, LXXILV, 6 (1977); John Searle, 'Minds, Brains and Programs', *The Behavioral and Brain Sciences* 3:3 (1980).

³ In explaining why he thinks computers *can* (or will someday), Marvin Minsky (in 'Why People Think Computers Can't', *AI Magazine*, Fall 1982), seems most impressed, for instance, with the fact that "computers can manipulate *symbols*."

⁴ John Searle, 'Minds, Brains and Programs', *The Behavioral and Brain Sciences*, 3:3 (1980); Ned Block, 'Troubles with Functionalism', in Wade Savage (ed.), *Perception and Cognition: Issues in the Foundations of Psychology*, Minnesota Studies in the Philosophy of Science, Vol. 9, Minneapolis, Minn.: 1978.

⁵ Paul Grice, 'Meaning', *Philosophical Review* 66 (1957), pp. 377–388.

⁶ See my *Knowledge and the Flow of Information* MIT/Badford: Cambridge, Mass., 1981.

WHAT'S IN A MIND?¹

INTRODUCTION

It has probably not escaped your notice that developing causal theories in psychology is incredibly difficult. Why should that be? Surely not for the sorts of reasons that are commonly given in introductory psychology classes. It's not that the science is young, for it is at least as old as physics. It's not that people are ever so complex, for so is the tiniest grain of sand. It's not just that organisms are much harder to experiment upon (although that is undeniably true), because that in part reflects the fact that we don't know how to ask the right questions – experimental methodology goes hand in hand with the development of theories. It seems to me that at least part of the reason that psychology is hard is that we don't have a good idea of what it's about – what it's a science *of*. For example, we are frequently told that psychology is the science of (human or animal) behavior. But that can't be true since physics does a very good job of *that*. For example, it predicts that if you jump off the top of a tall building your behavior will consist in accelerating at just under 10 meters per second every second. The same is true of more microscopic behaviors like those that relate the acoustical properties of the sounds you make to the way you move your lips and tongue. Clearly that's not the sense of behavior that's relevant to psychology. But what other sense of behavior do we intend when we characterize what psychology is about? Many people would be surprised to find that it isn't possible to give an answer to that question without implicitly taking a strong stand on a number of contentious philosophical issues.

Of course it is really asking too much that there be a clear conception of what psychology is about since, after all, the grouping of people and interests into a discipline is at least in part a political matter and nobody can dictate how it should go. Yet if there is to be a causal explanatory science within some domain, such groupings cannot be entirely insensitive to whether or not there are within that domain at least some *natural kinds*, or some clusters of phenomena

that fall under a uniform set of explanatory principles. This point is worth a brief digression.

NATURAL DOMAINS AND SCIENTIFIC EXPLANATION

It has been fairly commonplace to say, ever since Plato first said it, that we must endeavour to carve nature at her natural joints. But natural joints are not easy to find and they certainly cannot be specified in advance. There is no way of determining which phenomena will cluster together to form the object of a special science except by observing the success, or lack of it, of attempts to build theories in certain domains. But despite general concurrence on this point, people are nonetheless extremely reluctant to accept a partition of phenomena that leaves out something they feel is important – as if the development of a scientific field must be guided by a priori considerations of what matters to *us*.

Examples of the reluctance to part with a priori views of how phenomena must cluster can be found in studies of language, where the autonomy of phonology and syntax is still vigorously resisted, largely on the grounds that the cluster we call “language” includes much more than grammar. And of course language does indeed involve much more, but that does not mean that we will be able to develop a uniform explanatory theory of language phenomena when the latter are understood to include everything in the intuitively defined cluster. The same is true of vision. Recent successes in the study of visual perception (see Marr, 1982), depend on separating certain processes (called “early vision”) from the mass of knowledge-dependent processes that are ultimately involved in perception, i.e., from the entire perceptual process whereby optical stimulation eventuates in beliefs about the world.

But there are even more dramatic examples than these. For instance, in viewing psychology as the science of mental life, many people have been implicitly committed to the view that we should seek principles that apply over the domain of conscious experiences, such as J. S. Mill’s principles of “mental chemistry” that govern our conscious thoughts, images, impressions, perceptions, feelings, moods, and so on. Such an enterprise has met with little success. The reason most frequently given is methodological – it’s very difficult to

observe our experiences in an unobtrusive way. But there is another, perhaps even deeper reason for the failure. If the even modest success in building small scale theories in contemporary "information processing psychology" is any indication, we had no right to the *a priori* assumption that the set of conscious contents is a natural domain. Among the things we are conscious of are some that follow quasi-logical principles (common-sense reasoning), others that might follow associative principles (idle musings, mind wandering), and still others that may be governed primarily by biochemical principles (occurrences of pains, moods, or certain emotional states). Furthermore, it appears that in order to account for some regularities we need to posit processes and mental states that behave exactly like some states of consciousness, but of which we have not the slightest awareness. I consider it to be an empirical discovery of no small importance that if we draw the boundary around phenomena in such a way as to cut across the conscious-nonconscious distinction we find that we are at least able to formulate moderately successful mini-theories in the resulting cluster. The information processing literature is full of quite persuasive examples of such cases.

Now it is by no means obvious that in order to develop explanatory theories we should have to group together deliberate reasoning and problem solving processes with processes that seem to happen in a flash and with no apparent awareness that there is any mental activity going on, while at the same time to leave out of the picture such very vivid mental contents as the experience of pain, the experience of seeing a certain colour, the experience of being dizzy, or the experience of feeling the pieces of a puzzle fall into place. Not only is it not obvious, but it is vigorously opposed by many people who view this sort of regrouping as an indictment of information processing psychology. For example, in an otherwise insightful book on "visual thinking" Rudolph Arnheim (1969) attacks a certain computer model which solves visual analogy problems. His attack is based on the observation that when people solve such problems they go through a sequence of mental states involving a "rich and dazzling experience" of "instability", of "fleeting, elusive resemblances", and of being "suddenly struck by" certain perceptual relationships, whereas the program doggedly pursues a quasi-logical search, using pattern-matching and comparison processes. Similarly Hubert Dreyfus (1979) accuses the Artificial Intelligence community of not taking seriously

such distinctions as between focal and “fringe” consciousness and of ignoring our sense of oddness about the information processing postulates. And I can’t resist adding to this list Steve Kosslyn’s (e.g., Kosslyn, Pinker, Smith and Schwartz, 1979) accusation that I fail to take seriously subjects’ introspective reports in my analysis of mental imagery experiments. But more on this later.

The conscious-unconscious distinction is not the only one we may have to cut through in seeking a natural domain within the set of phenomena now covered by the term “psychology”. It may also turn out that large parts of such psychologically important domains as learning, development, emotion, mood, psychopathology, and psychomotor skills will not be explained in the same way as other kinds of psychological regularities. Phenomena in such domains may have to be partitioned in ways that are different from the way in which pretheoretical intuitions and everyday common sense draws the boundaries. We simply have no right to require that the categories of phenomena in our evolving science follow any particular *a priori* boundaries. Which phenomena will cluster into those explainable by some uniform set of principles is something we discover as we attempt to build theories beginning with the clear cases. Thus when John Haugeland (1978) disparages current information processing theorizing on the grounds that it cannot deal with the development and control of motor skills or with the pervasive effects of moods on cognitive activity, or when Paul Churchland (1980) does the same, citing the example of what he calls “large scale learning” or conceptual change, they are making presumptions about what ought to be in a certain natural scientific domain. In particular they are requiring that the role of moods, skills, and conceptual change be explainable within the same class of theory that would explain how people understand a sentence. We have no more right to set such *a priori* conditions on an evolving science than Galileo had of requiring *a priori* that planets follow the same laws of motion as cannonballs, or that the generative semanticists had of requiring *a priori* that all cases of the acceptable-unacceptable distinction among sentences be expressible within a common body of theory. It was an empirical discovery that planets did fall under the evolving laws of dynamics and that some distinctions among sentences were not part of a developing linguistic theory. In fact among those that were not, were some pretty familiar distinctions

of prescriptive grammar, such as when to use “between” and when to use “among”.

To return to the question I raised at the outset about why psychology is so hard the answer may very well be that there is no such natural domain. There may, instead, be a variety of different theoretical sciences that will stake out parts of the field. In addition there may also be a descriptive or taxonomic discipline which catalogues a variety of useful regularities and perhaps even ties them together with elaborate systems that serve more as mnemonic aids than as explanations – as I believe is the case with Freudian theory. And there could also be a number of useful engineering disciplines based on such systematic catalogues. I have always been impressed with how it is possible to do truly useful things in this world while being guided by an obviously false theory of why it works – behavior modification being a case in point, along with quite a few other theoretical bases for psychotherapy.

But I am not entirely persuaded of this pessimistic view, because I do see some reason to view recent developments as pointing to some better choices of taxonomies, ones that may be closer to picking out natural domains than we have had heretofore. I have already alluded to the example of work on language and to work in vision being carried out within the computational paradigm. Both have carved out relatively closed domains within which there is some hope that an explanatory theory will emerge. But what about the rest of the mind? Is there no hope that some useful distinctions can be made there as well?

The rest of the mind is indeed much harder. Everything there appears to be ever so holistic and interactive. Nonetheless it is not without some joints. I will tell you what is just about the most fundamental joint there is in all of psychology, and will suggest that unless due recognition is given to the distinction on which this joint rests there is little hope of even formulating the puzzles of learning and cognition so that they might some day yield to explanations. The distinction I have in mind is between two ways of attempting to explain an observed regularity in the behavior of some system. One way is by appealing to properties or mechanisms that are intrinsic to the system, or by appealing to certain functional capacities of the system (what I have elsewhere called its functional architecture). The

other is by assuming that the system has certain *representations* which it manipulates according to nomologically arbitrary rules (i.e., regularities that are not subsumed by some particular natural law, such as the regularities expressed by rules of grammar or rules of inference). In the latter case, moreover, in order to capture the regularities in the system's behavior we must mention the content of the representations, i.e., semantics becomes a relevant relationship.

We can think of these as being two distinct levels of explanation, though I have generally resisted the temptation to call them levels because the term "level" tends to be used rather loosely to refer to any convenient degree of abstraction at which one might choose to describe something. When I say that these two represent different levels of explanation I mean that there are two natural domains, or two potentially autonomous sciences based on the two disjoint vocabularies. It is therefore an empirical matter whether or not there really are two distinct levels. Having thus forewarned you as to my intentions, let me now give some simple examples to illustrate what I mean by the two levels. The examples may not be ideal, but at least they illustrate my point.

CONSTRAINTS AS DUE INTRINSIC PROPERTIES OF MECHANISMS OR TO REPRESENTATIONS

Suppose I showed you a black box (such as that illustrated in figure 1) into which I had inserted an electrode or some other response recorder. As we observe the box go about its usual function we discover that the ensuing record exhibits certain regularities. For example, we observe that either individual short pulses or pairs of such short pulses frequently occur in the record, and that when there are both pairs and single pulses (as sometimes happens) the pair appears to regularly precede the single pulse. After observing this pattern for some time we discover that there are occasional exceptions to this order but only when the whole pattern is preceded by a pair of long and short pulse sequences. Being scientists we are most interested in giving an explanation of this regularity. What kind of explanation will be most appropriate?

The answer, I maintain, depends upon what sort of device the black box is and in particular on what its *capacity* is beyond the particular behavior we have just been observing (i.e., not on what it is doing, or

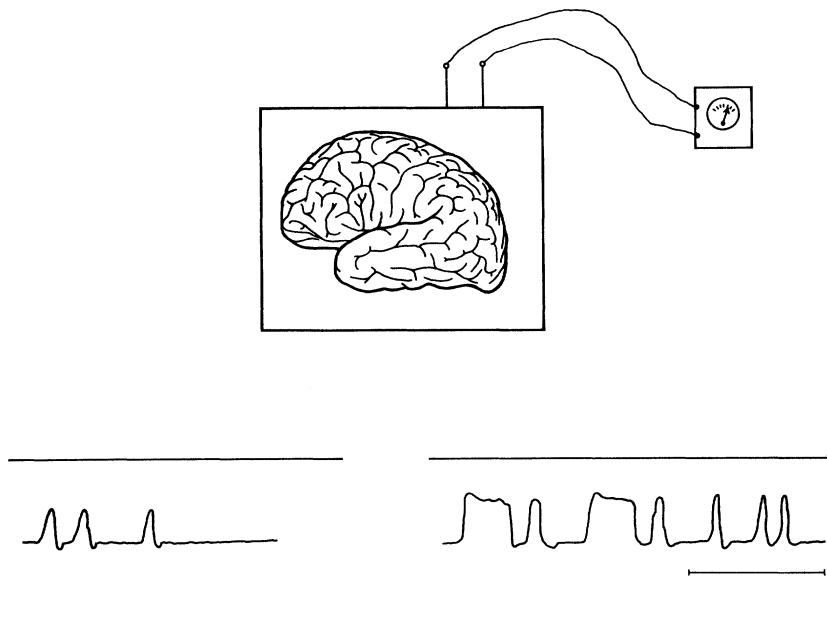


Fig. 1. Systematic patterns of behavior recorded from an unknown black box. The problem is to explain the observed regularity. (Reprinted, with permission, from Z. Pylyshyn: 1984, 'Computation and Cognition: Toward a Foundation for Cognitive Science', MIT Press/Bradford Books, Cambridge, Mass.)

what it typically does, but on what it *could* be doing in certain counterfactual situations). In this particular example, chosen deliberately to make a pedagogical point, I can confidently tell you that we would not find the explanation of its behavior in its internal structure, nor in any properties intrinsic to the box or its contents.

Now that might strike some people as an odd claim. How can the behavior of a system not be due to its internal construction or its inherent properties? What else could possibly explain the regularities it exhibits? In the end, of course, it is the existence of certain properties in the box that govern the totality of its behavioral repertoire, or its capacity. But as long as we have only sampled some limited scope of this repertoire (say, what it "typically" or "normally" does) we may not be in any position to infer what its intrinsically

constrained capacity is, hence the observed regularity may tell us nothing about the internal structure or inherent properties of the device.

Let us make this point more concrete by considering the question of why the black box in our example exhibits the particular regularity claimed. I can now reveal to you that the real reason the black box exhibits this regularity is simply that it is a box for transmitting (or, if you prefer, for "thinking in") English words encoded in International Morse Code. Thus the regularity we have discovered is attributable entirely to a spelling rule of English (viz., *i* before *e* except after *c*), together with the IMC code convention. And the reason that providing a detailed description of the component structure and the operation of the box would not explain this regularity is that the structure is capable of exhibiting a much greater range of behaviors – *the observed constraint on its behavior is not due to its intrinsic capability but to what its states represent*. I will later return to this idea that constraints on behavior can exist at different autonomous levels.

Let's now take another example, which will take us closer to a point about psychological explanation that I want to develop. Consider the regularities of colour mixing (e.g., perceived yellow light and perceived red light mix to produce perceived orange light). What sort of account might we expect as an explanation for these regularities? One which appeals to intrinsic properties of the system, to certain internal biological mechanisms, or one which (as in the Morse Code example) appeals to properties of *what is represented* rather than of the system itself? The question is an empirical one and I wish simply to point out what is at issue and on what kinds of empirical considerations the answer depends. In this case all the evidence points to there being a biological or biochemical mechanism responsible for the regularity. One of the reasons for expecting such an account (apart from the fact that we have quite a large fragment of the account already in hand) is the fact that the regularities appear to be largely insensitive to what subjects think they are looking at, to what they believe about the actual colour of these objects and the principles of colour mixing, and to what they think the purpose of the experiment is (within limits).

Contrast this with the case in which an investigator seeks to discover the principles of what might be called "imaginal colour mixing". Although I am not aware of this precise experiment having been reported in the literature, experiments quite similar to this exist in the

psychological literature (see the critical discussion in Pylyshyn, 1981). The experimenter asks subjects to imagine certain colours and to superimpose them in their mental image. The instructions might be something like this: "Imagine a transparent yellow filter beside a transparent red filter. Now imagine that the two filters are slowly moved together until they overlap. What colour do you see through the overlapped portion?" Suppose that the investigator discovers a set of reliable principles governing the mixing of imagined colours. What sort of explanatory account is likely to be the correct one in this case: an account based on appeal to biological or biochemical principles, or one based on what is being represented in the minds of the subjects – including what subjects tacitly know about the principles of colour mixing and what they take the task to be? Again it is an empirical question, though this time it seems much more likely that the "tacit knowledge" explanation will be the correct one. The reason for this is that it seems likely that the way colours mix in one's image will depend on what one knows about the regularities of perceptual colour mixing,² after all, we can make our image be whatever colour we want it to be!

The test for this explanation is to determine whether changing what the subject believes *by providing information* (possibly false information) will change the regularity in a logically explicable way. If it is, we say that the regularity is "cognitively penetrable" and conclude that no account based solely on appeal to intrinsic properties of a mechanism will by itself be adequate to explain the regularity or the way it can be altered. We draw this conclusion (just as we did in the Morse code example earlier), not because of an adherence to any dualist doctrine, but because we know that the evidence does not reveal a cognitively-fixed *capacity* inasmuch as the underlying mechanism is compatible with a wider range of behaviors than embodied in the empirically observed regularity. What the biological mechanism does provide is a way of *representing* or *encoding* the relevant knowledge, inference rules, decision procedures, and so on, not the observed regularity itself. (Incidentally, the mechanism also imposes certain resource limited constraints as well as providing what is called the control structure for accessing rules and representations, and these are what we have to appeal to in order to explain deviations from logical omniscience, but that is a topic for another occasion.)

I should emphasize that this is but a sketch of a position for which a

more complete argument is given in Pylyshyn (1984). Nonetheless it is a position which is implicit in much of contemporary cognitive science, and indeed much of psychology generally. For example, while one might consider searching for a neural or biochemical explanation for certain cognitively impenetrable psychophysical principles (e.g., the Weber function or the acoustical sensitivity curve or the Gestalt principles of perception), people would be surprised indeed if there was a biochemical or neurophysiological explanation of how we decide what the italicized pronoun refers to in the following sentences (taken from Winograd, 1972).

- The city councilors refused the demonstrators a permit because *they* feared violence.
- The city councilors refused the demonstrators a permit because *they* were communists.

Rather, we would expect the explanation to refer to one's knowledge of what city councilors are like, what the attitude of people in authority is to communists in certain countries, and so on. Only factors like this would explain why in particular cases the pronouns are assigned different referents in the two sentences and why the reference assignment could be easily changed by altering the context. For example, I recently inadvertently provided an example of the effect of context myself when I used these sentences in a talk I gave in Florence, Italy. I had forgotten that the city council of Florence was in fact drawn primarily from the communist party. Because of this my audience had assigned the same referent to the pronoun in both sentences and the point of the example had been largely lost!

There are at least two reasons that the appropriate explanation is one given in terms of beliefs, in spite of the fact that there can be no doubt that the behavior is *caused* by biological processes of some kind on each occasion. One reason is that it certainly need not be the case that the *same* neural or biochemical process is operative on every occasion where a pronoun is assigned a referent. This is just to repeat what I have already said before, namely that the category of behavior we are interested in may cross classify behavior described using biological terminology, so that a psychological generalization may collapse across an arbitrarily large disjunction of biologically distinct events.

The second reason is closely related, though somewhat harder to

state precisely. Because the regularity that governs which referent is assigned to which pronoun depends on the interpretation or meaning or semantic content that the listener assigns to the sentences, as well as to the meanings that are also assigned to other events and sentences (e.g., to the meaning you assigned to the sentence I used when I told you about the composition of the city council of Florence, which systematically altered the referent assignment process), we need to appeal to principles that are statable over semantic contents. Such principles include, for example, decision-theoretic ones, as well as principles of logic, of inference, of plausible reasoning, and so on. All such principles have something to do with the notion of a rational system. Although such a notion is not easy to explicate in a technically precise manner, it is implicit in all theorizing about cognition. We shall return to it briefly in the next section.

The idea that certain phenomena require knowledge-based explanations is not some eccentric view of mine but an extremely general view in practice. We never see attempts to provide neurophysiological explanations for why people assent to certain sentences or why they make certain sounds when asked where they live, or which style of clothes they prefer, and so on. The way we invariably explain such *interpreted* or meaningful behavior is in terms of processes that depend on the knowledge which subjects have and the way they use this knowledge to further some goal.

In spite of the ubiquity of such forms of explanation, it is not uncommon to assume that the use of notions like knowledge and goals is just a matter of convenience; that it is just easier and more practical for certain purposes to tell the story in terms of people's reasons and the decisions they make, but that in principle a more detailed neural explanation could be given if we only knew all the relevant facts. However, this is a misunderstanding of the nature of explanation. Any theoretical explanation presupposes a certain taxonomy under which the puzzling event is viewed. An explanation of a particular piece of emitted behavior viewed as a member of the category "generates a waveform with certain spectral properties" need not qualify as an explanation of *the very same event* viewed as a member of a category such as "claims that he is a liberal". The reason is quite simple: other members of the latter category can vary to an arbitrary degree in their acoustical properties. To explain the event under the latter taxonomy is to provide an account that applies to *all* members of that

equivalence class. An arbitrarily large disjunction of distinct neural or acoustical causal chains (one for each different way that a subject might have of making the claim) does not qualify, but an account that refers to the meaning of the sentence, to the subject's beliefs about politics, or to the subject's goals and intentions could (if the assumptions were independently motivated) do exactly that. And the reason that we want to view the event in terms of *that* category is simply that there are generalizations that are statable over that category that are not statable over other categories.

There are two general reasons why we need to postulate representations. The first is the plasticity and stimulus independence of meaningful behavior (i.e., interpreted behavior, rather than movements). If we view physics as providing constraints on states of the universe that are physically possible, then comparable constraints on what is psychologically possible have rarely been articulated, at least not in terms of S-R relations. If we ask what action is possible for a subject in a certain laboratory situation (e.g., what the subject is *capable* of doing or saying) there is very little that can be said outside of certain psychophysical constraints on perceptual-motor performance. That's because almost any prediction could be falsified if the subject had certain beliefs, and almost any relevant beliefs we can think of are within a human subject's cognitive capacity. The second reason is closely related. Whatever regularities in behavior we might observe under some set of conditions can be altered in a systematic and logically coherent way (to a first approximation) by merely providing the subject with certain information – by telling or showing the subject something, or providing cues which together with other beliefs warrant some plausible inference.

WHY IS FUNCTIONALISM NOT ENOUGH?

Before proceeding to elaborate on these two points I need to make a small aside. There are certain critics of the view I am advocating (e.g., Stich, 1984) who accept most of the above points, say about the stimulus independence of behavior and about the impossibility of capturing the right generalizations in a physical or biological vocabulary, but nonetheless insist that this can all be done in a purely functional theory. In other words they go along with the story I have been telling except that they don't see the necessity of talking about states

that have representational content. They see the need for only two levels of description – the physical and the functional – and feel that the semantic level does not introduce any new generalizations. This is a tempting view, since it skirts some hard problems of semantics. Yet there are a number of reasons why I am convinced that it cannot be sustained.

The general issue of meaning and semantics is an extremely difficult one that has occupied philosophy of mind for much of this century. I don't intend to raise the full spectre of this problem. All I want to do here is to briefly sketch two reasons why I believe that talking about functional states without mentioning what they represent will not do for the purpose of providing explanatory adequacy in psychology. Or rather, it will not do *all* that needs to be done. The first is that there is no reason to think that states individuated functionally are of the same grain as those individuated semantically or, to put it another way, some generalizations need to be stated over categories (such as “believes that the building is on fire”) that are not reducible to a finite disjunction of functional states. The second is closely related to our earlier discussion of capacities and I shall return to it presently.

To make the first point more precise we need to have a clearer idea of what is entailed in claiming that something is a functional explanation on the one hand, or that it is a semantic or representation-governed explanation on the other. Defining the notion of functional explanation in a technically precise manner is difficult because of certain problems about borderline or intuitively unclear cases (Block, 1978). Nonetheless the basic idea is quite straightforward: Any intrinsic property of a system that can be expressed without reference to the physical structure or composition of the system, and which captures regularities in its overt behavior is a functional property. In that respect the sorts of representation-dependent regularities we discussed earlier might be viewed as arising from some functional property or other of the system. And this is exactly right: just as there can be no behavior without a physical cause so there can be no behavior without a functional cause. But from this it does not follow that the generalizations concerning *patterns* of behavior can be stated in a functional vocabulary, any more than they can be stated in a physical vocabulary, and for much the same reason.

The reason that we cannot state some of the generalizations in a functional vocabulary, is that what counts as a case of the same

behavioral regularity, from the point of view of a representational (or semantic, or intentional) description, can arise from different functional causes. Since they are distinct behaviors, they can be distinguished by functional criteria. They may, for example, consist in the utterance of quite different sentences or in the execution of quite different behavioral acts. Yet they may nonetheless fall under a common semantic generalization such as, "Whenever S wants X to happen (e.g., X = Someone gives S money) and believes that it will only happen if he does A (e.g., A = come to work), then he will tend to do A." So long as there are generalizations statable over such categories we will have to find a vocabulary for expressing them.

One reason that such generalizations don't count as functional is that functional distinctions must be stated without regard to properties of the world that do not enter into a direct causal relation with the system whose behavior we are explaining. This is precisely what we have to do in the case of representational explanations: We have to say things like S did something or other *because* S had the belief (or goal) that P, or S represented something as R, where the contents of P and R are not properties that themselves cause the behavior. For example, if I run because I believe that I am being pursued by ghosts, it is not some ghost that causes my running. Nonetheless I have to mention ghosts in the explanation, otherwise there are many things about the pattern of my behavior that I would not be able to explain, such as the things that I say, the kind of information that will systematically affect my running behavior, and so on.

That there are generalizations statable in one vocabulary (e.g., a semantic or intentional one) that are not statable in another vocabulary (e.g., a functional one) should not in itself be surprising, since that is precisely the case between the functional and the physical levels of description. Different vocabularies pick out different equivalence classes of properties or behaviors, and therefore are capable of describing or explaining different patterns of regularities. For example, two systems that are in the same *intentional* or representational state (have the identical goals and beliefs) will behave in ways that fall under the same semantic generalizations. Yet they need not behave in ways that are identical in all respects. This can be made clear if we consider the case of a computer model of some cognitive process. The functional states of such a model correspond to the computational states of the system. These states determine the system's behavior.

Such a state can be specified by describing all the datastructures (viewed as syntactic objects) and by giving the state of execution reached by the program. Yet not all these properties have semantic interpretations in the cognitive domain.³ Not all components of such states are interpreted as goals, beliefs, or other “propositional attitudes”. Many properties that determine which computational or functional state the system is in, and hence which determine the observed behavior of the system – including such things as the relative amounts of time taken to carry out particular tasks and other measures of the task complexity – are not semantically interpreted properties of the system. These include properties of what I have called the functional architecture of the system, as well as such other computational properties as those that go under the heading of “control structures”. They are properties which specify, for example, when and how representations and rules will be accessed and which also determine various resource limits of the system.

Consequently, if we individuate functional states in the usual way, according to whether they lead to potentially observable differences in behavior, and if we individuate representations according to their semantic content, we shall find that the two correspond to quite different grains or levels of aggregation. Indeed we will find that there will in general be an unbounded number of functional states corresponding to a system having some particular belief or goal, just as there are an unbounded number of physical states corresponding in general to some particular functional state. Moreover, we need *both* levels of aggregation; the semantic level to capture meaning-dependent regularities (such as logical or rational relations), and the functional (or syntactic, or computational) to capture such things as reaction times and departures from ideal rational behavior.

The second reason I think there really is such a thing as an autonomous representational level, or as Allen Newell (1982) calls it, a “knowledge level” is related to the notion of constraints or *capacities* that I discussed earlier in connection with the Morse code example. One of the hallmarks of a true level is that it represents a set of constraining principles beyond the constraints imposed at a lower level and not derivable from principles at the lower level. This is true, for example, in going from phonology to syntax. But perhaps a better example might be what happens when we go from physics or chemistry to biology.

For example, assume that biology represents an autonomous level (i.e., that it picks out a natural scientific domain). Then there are principles of biology, including genetics, embryology, and so forth, not derivable from the laws of physics, that we need to appeal to in order to explain the behavior of objects that fall within the domain of biology, for example, principles of cellular growth and division, principles of metabolic conversion, of genetics, and so on, that apply to all living things. But physical laws are compatible with the existence of a much wider range of possible configurations of atoms or of carbohydrate and protein molecules than found in all known living creatures, and very likely a much wider range than could occur in any biologically *possible* living creature. How then can we explain why such other systems don't exist, given that they are compatible with the laws of physics?

Many explanations might be given. Note, however, that a distinguishing characteristic of all the existing objects in that domain is that certain of their properties (viz, the biological ones) can be explained in terms of biological principles. Such principles can explain, among other things, why these objects (i.e., living things) have to have certain properties in order to survive, to procreate, or in general to continue to exist as members of the natural kind of which they are members. Consequently an explanation of the nonoccurrence of certain kinds of systems (viz., ones with the wrong combination of biological properties) will thus also have to advert to such biological principles. For example, among the kinds of explanations that might be offered are such things as that some of these systems simply do not conform to certain biological requirements: perhaps they lack the mechanisms for reproduction, or they have certain characteristics that cannot be transmitted genetically, or they do not metabolize the chemicals in our environment, and so on. Each of these reasons refers to a principle that is statable only in the vocabulary of biology (again, assuming for the sake of argument that it is an autonomous science) and each explains something about why possibilities allowed at a lower level are not, as a matter of fact, actualized.⁴

Now it is my contention that exactly the same thing happens in going from the functional to the representational level, providing that we have a clear enough idea of what the functional level is. The clearest idea of a functional level comes from computational models. In a computational model certain fixed constraints are imposed by

what is called the functional architecture of the system. It specifies that only certain kinds of functions can be realized with certain sorts of time and memory complexity profiles. There are few concrete proposals for such a cognitive functional architecture at present. The ones there are typically include hypotheses about the size of a workspace in which symbols can be passed from process to process, hypotheses about how memory must be organized and accessed, constraints on how rules are invoked, and so on. Since they are functional, such constraints cannot mention content, only formal structural properties or computational mechanisms.

But now we notice that some sequences of functional states that are compatible with the given functional constraints never occur. We would like to give a principled account of this, just as Shimon Ullman (1979) set out to give an account of why certain logically possible mappings between two dimensional proximal stimuli and possible three dimensional scenes don't occur in vision: a pursuit that led him to the notion of a "natural constraint" that functions like a built-in assumption in the visual system. If we are allowed to talk about the states as representations, and if we are allowed to use semantic notions like validity, or to cite principles like rationality, we can give a natural account of the difference between those sequences of states that occur and those that don't. That's because some of the sequences would come out under our semantic interpretation as just crazy. We might get something like when the system believes that there is a crow on the roof it goes into the state of believing that supply side economics is working. We can exclude such a sequence as being semantically anomalous, but there is nothing in *functional* constraints in general to distinguish between crazy sequences and rational ones. On the reasonable assumption that most actual sequences of functional states in humans are, if not intelligent, at least not semantically anomalous and incoherent (rationality is the unmarked case), such semantic principles would serve exactly the same role in providing a principled distinction between the permissible and the realizable that was played by, say, genetic or embryological principles in the case of the biological example mentioned earlier. When we have such principled constraints, statable over a distinguished (in this case nonfunctional) vocabulary, this is a *prima facie* indicant of a separate level.

Notice that although many semantic relations can be mirrored in functional or syntactic relations – a fundamental discovery that is

responsible for the success of both computing and formal logic – not all semantic notions can be mapped onto syntax. In particular the difference between truth-preserving transformations and non-truth-preserving ones is not itself a syntactic one. Both types of transformations are expressible as formal rules. Indeed, both kinds are studied formally in the mathematical theory of rewriting systems (e.g., in the study of Post Production Systems, Lindenmeyer Systems, algebraic concatenation systems of various kinds, etc). In Logic, however, the *semantic* distinction between a *proof* and some other (invalid) set of transformations is fundamental. Similarly, in computer science semantic distinctions are also fundamental inasmuch as real computing has to do with the manipulation of symbols that have intended interpretations. Hence, notions of truth and correctness are important. These notions, in turn, cannot be defined syntactically any more than the notion of validity can in Logic (though it can sometimes be mirrored in syntactic rules – that's what makes mathematical logic possible).

That's probably more than I needed to say on the subject of content for the present purposes. The basic idea is that in order to explain some kinds of regularities, specifically regularities among interpreted categories of inputs and outputs (i.e., interpretations of perceptual events and interpreted actions), we need to refer to what internal states represent. And this, in turn arises from the fact that, as in the Morse code example, regularities are attributable at least in part to what the organism believes about some domain rather than to the organism's intrinsic capacities. Despite the familiarity of the basic thesis, the implications of this view are far reaching. Consider, for example, the implications of this distinction for an understanding of learning and development

LEARNING VERSUS OTHER KINDS OF ENVIRONMENTALLY-INDUCED CHANGES

The view I have been sketching suggests that it is a mistake to take a monolithic view of cognitive change, as has been routine in most empiricist psychology and much of the early work in cybernetics aimed at developing "self-adaptive systems". The reason is that such change can be of two very different kinds. One type of change arises as a rational consequence of certain information-bearing events which

the organism experiences. The organism forms a cognitive *representation* as a result of such events and this allows it to apply rules to these representations and hence to infer certain beliefs about the world, to draw nondemonstrative inferences or to induce plausible hypotheses. This type of change is what has always been known in the vernacular as "learning".

It is important to distinguish this type of effect from other ways of producing changes in an organism, including changes due to variations in nourishment, changes due to growth and maturation of glands and organs (including the brain), changes due to injury or trauma, and perhaps even changes due to noninformative aspects of reinforcement and practice, i.e., effects over and above those that arise because the reinforcers serve as signals which inform subjects of the contingencies and hence allow them to adjust their utilities accordingly (assuming there is such a thing as a noninformative aspect of reinforcers – see Brewer, 1974).

The reason we must distinguish two different kinds of relations between organisms and environments is the same as the reason that we had to distinguish two different levels of description of processes in our earlier discussion, that is, that the two follow quite different principles. For example, rational fixation of belief arising from environmental information, unlike just any arbitrary changes in biological state caused by the environment, can proceed in radically different ways depending upon how the organism *interprets* the stimulation, which in turn depends on what prior beliefs and utilities it has.

Now it is the knowledge-acquisition type of effect that concerns educators and which corresponds to the everyday notion of learning as "finding out about" something or other, or discovering something to be the case, or finding out how to do something. At the same time, however, the appropriateness of *this* kind of process as an account of how certain states of an organism are arrived at or how certain capacities are achieved has increasingly come under question in recent years. More and more frequently it has been argued that the information in an organism's environment is too impoverished to allow it to logically *infer* certain representations that form part of its mature state. This argument has been made with particular force in the case of language by Chomsky and by others interested in the issue of language learnability (e.g., Wexler and Culicover, 1980).

Such critiques of the learning view have frequently been misunder-

stood as implying that the final cognitive competence (say for language or conceptualization) is already present at birth, or that no influence of the environment is necessary for its development. But that is not the case. Environmental stimulation, sometimes of a very specific sort, is undeniably necessary for the achievement of most cognitive competences. However, that is not the issue that concerns some of the arguments about learning. The stimulation required to produce in me the belief that this building is on fire is extremely specific to that belief. Stimulations in that category have virtually no chance of causing me to have the belief that the sky is falling, or any other logically unrelated belief. Yet the set of events that have my belief that the building is on fire as their consequence are only specific in an informational or semantic sense. They need have nothing in common from the point of view of their physical form: I could hear a shout, be handed a note, see flames, smell smoke, hear a bell which I *interpret* to be a fire alarm, or even hear a tapping whose pattern corresponds to some prearranged arbitrary code. Not only is this set of events open-ended but it can be systematically changed by the provision of other ancillary information, as in the case of prearranging the code convention, or someone telling me that the fire alarm bell is being tested.

The point is that the mere fact of specific environmental cause is no more reason to believe that the attained cognitive state is learned than is the fact of the influence of sunlight on skin colour reason to believe that suntans are learned. Something more is required to show that learning is involved. What is needed is a demonstration that the cognitive state is arrived at because the environmental conditions provided the organism with certain information about things in the world, information from which the organism is able to infer certain beliefs whose content is related to those conditions.

Now if that is what is needed it should be straightforward, at least in principle, to inquire whether the specific environmental conditions which do lead to the cognitive state in question provide sufficient information (or adequate premises) *in principle* for the logical construction or induction of that state. If the answer is "yes" then learning is certainly a plausible hypothesis. If, on the other hand, the answer is something like "Yes, but only if the organism makes the following additional assumptions, or if for some reason it is prohibited from considering the following alternative hypotheses . . .", then we know at

least that the state is not attained solely by learning, at least not from the conditions in question. In such cases *something*, such as the "assumptions" or the constraints that prohibit certain hypotheses, must have been achieved by methods other than learning.⁵ This is the burden of the research by people like Wexler and Cullicover (1980). Such constraints or "assumptions" must, in other words, be part of what I referred to as the functional architecture. Consequently they must either be part of the initial state, or have developed from the initial state (with or without environmental influence), by noncognitive means (where by noncognitive I mean by processes that do not operate on the basis of rules and representations; are not inferential or otherwise knowledge-dependent). This is what people mean, or should mean, when they claim that some cognitive state or capacity is *innate*: not that it is independent of environmental influence, but that it is independent of rule governed construction from representations of relevant properties of the environment. In other words, it is not systematically related to its environmental cause solely in terms of its semantic (or, informational) content.

A closely related distinction arises in connection with research on vision by people at MIT and elsewhere. Here it has been demonstrated repeatedly that there is insufficient information in the light not only to warrant the perceptual beliefs that the light gives rise to, but also to unequivocally determine the knowledge-independent low level visual representations we form of even unfamiliar scenes. However, it has also been shown by people like Shimon Ullman (1979), that if certain assumptions about the nature of the distal environment are made then a unique mapping from proximal stimulus to such representations is possible. Such assumptions include, for example, that visually interpretable scenes consist only of rigid three dimensional objects, or that the surfaces in the scene are almost everywhere continuous, or that the reflectance of surfaces is nearly Lambertian. These assumptions cannot be *inferred* from the light emanating from the scene. They must come from elsewhere, in particular they must be internal to the visual system, since we know that the relevant part of the visual system is cognitively impenetrable. Furthermore, if such assumptions are not inferred from other sources of information – and there is persuasive evidence that at least some of them are not (e.g., 2 week old children appear able to distinguish surfaces differing only in their distance away from the child) – then they must be part of the initial state or develop

from the initial state by noncognitive means. They thus appear to have the status of Kantian transcendental categories.

Before concluding this sketch of the representational view of the mind, let me turn briefly to a slight rewording of a question that Shakespeare once asked (and Warren McCulloch adopted as a title for a famous paper), “What’s in the mind that ink may character?” I have suggested that what’s in the mind is *representations*. Now that in itself should not come as a revelation unless you have already lost your innocence through behaviorism or some other esoteric schooling. After all, we all know that what determines what we do is what we know and want, and what in the end we decide to do. Well, then, if that’s all there is to it why is psychology so hard? And why is it that I seem not to know my own mind even though I appear to have access, at least some of the time, to my thoughts? For example, I know that some of my thoughts are in the form of words while others are in the form of images, that some are vivid while others are vague or fleeting, that some of my thoughts appear to be deliberate while others unfold as though they had a life of their own. When I introspect I notice that I have a reasonably accurate picture of the world and that objects in my images generally tend to obey the laws of physics, as if there were a Humean preestablished harmony. Why can I not put together some careful observations of this sort and build an explanatory theory of mental function?

The answer, I believe, is that we are seriously deluding ourselves when we think that introspection gives us access to psychological processes or to properties of mental representations. What we have access to is the *content* of our mental states, to what our mental states are *about*, roughly speaking. Our introspections may tell us that our thoughts are *about* tables and chairs and rooms of people, about sailboats and Nobel prizes, about money and fame and other unattainables or intangibles. None of these things are themselves mental states. Indeed, most of them probably don’t even exist. They are states of what logicians sometimes call “possible worlds”. It is a mistake to speak of these contents as literally in the mind or the head since only representations of them are. We can’t describe our actual mental processes because we literally don’t have the words for them. Our words are meant to describe *things*, not their mental counterparts.

The point then, is that our mental states are about *things* and that we only have access to the content of our mental states. Consequently

the properties we report in experiments on imagery are not properties of representations, but properties that we represent the imaginary *things* as having. This simple and, one would have thought obvious, point is quite a deep one that is misunderstood by almost every psychologist who has theorized about mental imagery. If we could find out all we ever wanted to know about an organism's mental experience we would still have a long way to go to develop a psychological theory, since what we would find out is what properties people believe the things they are imagining actually have, not properties that their images have.

When we study what subjects do with their mental images we are primarily studying what properties these subjects believe that things they *could look at* would have. Thus when psychologists ask subjects to *imagine* that they are visually scanning a map and find that the time to scan increases with distance, they are showing that people *believe* that when you visually scan a map it takes longer to scan greater distances. Similarly, when they ask subjects to imagine looking at a small object they find that it takes longer for them to detect the presence of some visual feature than it does when they imagine that they are looking at a large object. In this case it is a reasonably safe bet that it will turn out on closer scrutiny that what it shows is that subjects *know* that when you look at smaller objects the details are harder to see so it would probably take longer in that case. We have conducted experiments that I believe show this quite conclusively, at least for certain cases. What these experiments show is that when subjects do not interpret their task as imagining that they are actually *seeing some real event take place* (and hence using their knowledge to infer what would happen) then these sorts of results can be made to go away. I will not belabor this point because you can all read the argument in Pylyshyn (1981).

The more interesting question with respect to such research on imagery seems to me to be why ever would anyone expect that what's in the mind has *any* of the properties that the things we are thinking about have (like spatial extent, rigidity, or conformance to Newtonian dynamics)? And the answer, it seems to me, goes back to the extremely seductive feeling we have that when we close our eyes we are examining the contents of our mind. This seduction is reinforced by the equally strong impression that we are not doing any actual *reasoning* at the time because it's all so quick and easy for us.

But as Dan Dennett (1978) once said about the mind, it's really quite cold and dark in there. And, anyway, who knows what it feels like to reason? What does it feel like to understand a sentence like the ones I showed earlier that indisputably involve reasoning? The only way you will ever know what goes on in the mind is by using the same kind of methods that chemists and solid state physicists use, that is, by building models of those aspects that constitute a natural domain and then, when you find that they explain all the facts in that domain, you take those models to be *true*, to be a *literal* description of reality. You do this regardless of how strange the resulting picture might appear to the naive observer, with its arrays of impersonal formal symbols. Psychologists used to complain that the mind could not possibly contain NPs and VPs as the linguists supposed, and indeed took this as an argument against the relevance of formal linguistics. But building models that meet empirical boundary conditions is the only way anyone ever finds out what some unobservable phenomenon is made of. There is no other way of finding out what's *really* there. Indeed, in science that's the only sense of reality there is.

NOTES

¹ Loosely based on my presidential address, 'Whereof One Cannot Speak . . .' delivered to the Eighth Annual Meeting of the Society for Philosophy and Psychology (London, Ontario, May 1982). Unfortunately, the jokes had to be omitted in this version. For reprints write the author at the Department of Psychology, University of Western Ontario, London, Ontario, Canada N6A 5C2.

² Note that one needn't always use what one knows in doing some particular task. Thus, for example, if the experiment described above is carried out people frequently do such things as free associate to the colour names, guess, or do just about anything depending on how the task is presented to them. Nothing much follows from this concerning the nature of their imagery mechanism. This can be easily be seen by observing that the same thing happens when the experiment involves imaging a sequence of a number, a plus sign, another number and an equal sign, and the instructions are to imagine the number that comes next. Here too subjects can easily refrain from using their knowledge of arithmetic and imagine any number. But that, of course, tells us nothing about the mechanisms of imagery either: making valid inferences is not the only thing we can do with our knowledge!

³ Note that I speak of some of the datastructures, etc., as "having" semantic interpretations without at this stage prejudging the question of why they have such interpretations – and whether the interpretation or the semantics is internal to the system or external. In fact the story that I am telling so far is compatible with a "methodological solipsism" view (Fodor, 1980) which says that the meanings of the symbols is not in the

machine at all. By the same token, the story is also compatible with other views of the nature of the semantics assigned to the symbols – including a “conceptual role” view or some sort of causal view (e.g., Dretske, 1981). For the time being I will not be concerned with where the interpretation comes from and where it resides. All I want to claim is that an explanation of behavioral regularities must advert to such semantic content. Moreover, talk about semantic content is not eliminable in favour of talk of functional states because, at the very least, a particular semantic content picks out an equivalence class of functional states and neither the classes nor the principles governing their regularities can be characterized in a functional vocabulary – just as functional states are not eliminable in favour of physical states because they correspond to equivalence classes of physical states that cannot be characterized in a physical vocabulary.

⁴ Note by the way, that the constraint is a relative one: it says that you could not have a system with such and such properties which was at the same time a member of domain D and thus subject to the principles of domain D, not that a system with those properties could not exist at all. The point is not that physics is too permissive, only that there are constraints on the realization of physically possible configurations that cannot be *explained* within the resources of the vocabulary and principles of physics, just as there are various *generalizations* that cannot be captured within these same resources.

⁵ I intend the phrase “specific environmental conditions that lead to the state in question” to refer to the entire set of conditions that are necessary for the final state to be attained. Thus if one argues that the enabling “assumptions” referred to here are themselves learned it would simply indicate that we had not provided a sufficiently broad set of conditions for the learning in question. Sometimes, as in the case of learning grammars, we do have some idea of what the relevant set of conditions are like. Thus people who work on the problem of language learnability have shown, with only weak assumptions on the boundary conditions for language acquisition, that grammars could not be learned without strong constraints on which grammars the organism could entertain as candidates.

REFERENCES

- Arnheim, R.: 1969, *Visual Thinking*, University of California Press, Berkeley.
- Block, N.: 1978, ‘Troubles with Functionalism’, in C. W. Savage (ed.), *Perception and Cognition: Issues in the Foundations of Psychology (Minnesota Studies in the Philosophy of Science, Vol. IX)*, University of Minnesota Press, Minneapolis.
- Brewer, W. F.: 1974, ‘There is No Convincing Evidence of Operant or Classical Conditioning in Adult Humans’, in W. B. Weimer and D. S. Palermo (eds.), *Cognition and the Verbal Processes*, Prentice-Hall, Hillsdale.
- Churchland, P. M.: 1980, ‘Plasticity: Conceptual and Neuronal’, *Behavioral and Brain Sciences* 3, 133–134.
- Dennett, D.: 1978, ‘Toward a Cognitive Theory of Consciousness’, in C. W. Savage (ed.), *Perception and Cognition: Issues in the Foundations of Psychology (Minnesota Studies in the Philosophy of Science, Vol. IX)*, University of Minnesota Press, Minneapolis.
- Dretske, F.: 1981, *Knowledge and the Flow of Information*, Bradford/MIT Press, Cambridge, Mass.

- Dreyfus, H. L.: 1979, *What Computers Can't Do*, Harper and Row, New York.
- Fodor, J. A.: 1980, 'Methodological Solipsism Considered as a Research Strategy for Cognitive Psychology', *Behavioral and Brain Sciences* **3**, 63-73.
- Haugeland, J.: 1978, 'The Nature and Plausibility of Cognitivism', *Behavioural and Brain Sciences* **2**, 215-260.
- Kosslyn, S. M., S. Pinker, G. Smith, and S. P. Shwartz: 1979, 'On the Demystification of Mental Imagery', *Behavioral and Brain Sciences* **2**, 535-581.
- Marr, D.: 1982, *Vision*, W. H. Freeman, San Francisco.
- Newell, A.: 1982, 'The Knowledge Level', *Artificial Intelligence* **18**, 87-127.
- Pylyshyn, Z. W.: 1981, 'The Imagery Debate: Analog Media Versus Tacit Knowledge', *Psychological Review* **88**, 16-45.
- Pylyshyn, Z. W.: 1984, *Computation and Cognition: Toward a Foundation for Cognitive Science*, Bradford/MIT Press, Cambridge.
- Stich, S.: 1984, *Folk Psychology and Cognitive Science: The Case Against Belief*, Bradford/MIT Press, Cambridge, Mass.
- Ullman, S.: 1979, *The Interpretation of Visual Motion*, MIT Press, Cambridge.
- Wexler, K. and P. Culicover: 1980, *Formal Principles of Language Acquisition*, MIT Press, Cambridge, Mass.
- Winograd, T.: 1972, 'Understanding Natural Language', *Cognitive Psychology* **3**, 1-191.

Departments of Psychology
and Computer Science
University of Western Ontario
London, Ontario N6A 3K7
Canada

PART II

CONNECTIONIST CONCEPTIONS

CONNECTIONISM, ELIMINATIVISM, AND THE FUTURE OF FOLK PSYCHOLOGY¹

1. INTRODUCTION

In the years since the publication of Thomas Kuhn's *Structure of Scientific Revolutions*, the term 'scientific revolution' has been used with increasing frequency in discussions of scientific change, and the magnitude required of an innovation before someone or other is tempted to call it a revolution has diminished alarmingly. Our thesis in this paper is that if a certain family of connectionist hypotheses turn out to be right, they will surely count as revolutionary, even on stringent pre-Kuhnian standards. There is no question that connectionism has already brought about major changes in the way many cognitive scientists conceive of cognition. However, as we see it, what makes certain kinds of connectionist models genuinely revolutionary is the support they lend to a thoroughgoing eliminativism about some of the central posits of common sense (or 'folk') psychology. Our focus in this paper will be on beliefs or propositional memories, though the argument generalizes straightforwardly to all the other propositional attitudes. If we are right, the consequences of this kind of connectionism extend well beyond the confines of cognitive science, since these models, if successful, will require a radical reorientation in the way we think about ourselves.

Here is a quick preview of what is to come. Section 2 gives a brief account of what eliminativism claims, and sketches a pair of premises that eliminativist arguments typically require. Section 3 says a bit about how we conceive of common sense psychology, and the propositional attitudes that it posits. It also illustrates one sort of psychological model that exploits and builds upon the posits of folk psychology. Section 4 is devoted to connectionism. Models that have been called 'connectionist' form a fuzzy and heterogeneous set whose members often share little more than a vague family resemblance. However, our argument linking connectionism to eliminativism will work only for a restricted domain of connectionist models, interpreted in a particular way; the main job of Section 4 is to say what that domain is and how the models in the domain are to be interpreted. In Section 5 we will illustrate what a connectionist model of belief that comports with our strictures might look like, and go on to argue that if models of this sort are

correct, then things look bad for common sense psychology. Section 6 assembles some objections and replies. The final section is a brief conclusion.

Before plunging in we should emphasize that the thesis we propose to defend is a *conditional* claim: *If* connectionist hypotheses of the sort we will sketch turn out to be right, so too will eliminativism about propositional attitudes. Since our goal is only to show how connectionism and eliminativism are related, we will make no effort to argue for the truth or falsity of either doctrine. In particular, we will offer no argument in favor of the version of connectionism required in the antecedent of our conditional. Indeed our view is that it is early days yet – too early to tell with any assurance how well this family of connectionist hypotheses will fare. Those who are more confident of connectionism may, of course, invoke our conditional as part of a larger argument for doing away with the propositional attitudes.² But, as John Haugeland once remarked, one man's *ponens* is another man's *tollens*. And those who take eliminativism about propositional attitudes to be preposterous or unthinkable may well view our arguments as part of a larger case against connectionism. Thus, we'd not be at all surprised if trenchant critics of connectionism, like Fodor and Pylyshyn, found both our conditional and the argument for it to be quite congenial.³

2. ELIMINATIVISM AND FOLK PSYCHOLOGY

‘Eliminativism’, as we shall use the term, is a fancy name for a simple thesis. It is the claim that some category of entities, processes or properties exploited in a common sense or scientific account of the world do not exist. So construed, we are all eliminativists about many sorts of things. In the domain of folk theory, witches are the standard example. Once upon a time witches were widely believed to be responsible for various local calamities. But people gradually became convinced that there are better explanations for most of the events in which witches had been implicated. There being no explanatory work for witches to do, sensible people concluded that there were no such things. In the scientific domain, phlogiston, caloric fluid and the luminiferous ether are the parade cases for eliminativism. Each was invoked by serious scientists pursuing sophisticated research programs. But in each case the program ran aground in a major way, and the theories in which the entities were invoked were replaced by successor theories in which the entities played no role. The scientific community gradually came to recognize that phlogiston and the rest do not exist.

As these examples suggest, a central step in an eliminativist argument will typically be the demonstration that the theory in which certain putative entities or processes are invoked should be rejected and replaced by a better theory. And that raises the question of how we go about showing that one theory is better than another. Notoriously, this question is easier to ask than to answer. However, it would be pretty widely agreed that if a new theory provides more accurate predictions and better explanations than an old one, and does so over a broader range of phenomena, and if the new theory comports as well or better with well established theories in neighboring domains, then there is good reason to think that the old theory is inferior, and that the new one is to be preferred. This is hardly a complete account of the conditions under which one theory is to be preferred to another, though for our purposes it will suffice.

But merely showing that a theory in which a class of entities plays a role is inferior to a successor theory plainly is not sufficient to show that the entities do not exist. Often a more appropriate conclusion is that the rejected theory was wrong, perhaps seriously wrong, about some of the properties of the entities in its domain, or about the laws governing those entities, and that the new theory gives us a more accurate account of *those very same entities*. Thus, for example, pre-Copernican astronomy was very wrong about the nature of the planets and the laws governing their movement. But it would be something of a joke to suggest that Copernicus and Galileo showed that the planets Ptolemy spoke of do not exist.⁴

In other cases the right thing to conclude is that the posits of the old theory are reducible to those of the new. Standard examples here include the reduction of temperature to mean molecular kinetic energy, the reduction of sound to wave motion in the medium, and the reduction of genes to sequences of polynucleotide bases.⁵ Given our current concerns, the lesson to be learned from these cases is that even if the common sense theory in which propositional attitudes find their home is replaced by a better theory, that would not be enough to show that the posits of the common sense theory do not exist.

What more would be needed? What is it that distinguishes cases like phlogiston and caloric, on the one hand, from cases like genes or the planets on the other? Or, to ask the question in a rather different way, what made phlogiston and caloric candidates for elimination? Why wasn't it concluded that phlogiston is oxygen, that caloric is kinetic energy, and that the earlier theories had just been rather badly mistaken about some of the properties of phlogiston and caloric?

Let us introduce a bit of terminology. We will call theory changes in which the entities and processes of the old theory are retained or reduced to those of the new one *ontologically conservative* theory changes. Theory changes that are not ontologically conservative we will call *ontologically radical*. Given this terminology, the question we are asking is how to distinguish ontologically conservative theory changes from ontologically radical ones.

Once again, this is a question that is easier to ask than to answer. There is, in the philosophy of science literature, nothing that even comes close to a plausible and fully general account of when theory change sustains an eliminativist conclusion and when it does not. In the absence of a principled way of deciding when ontological elimination is in order, the best we can do is to look at the posits of the old theory – the ones that are at risk of elimination – and ask whether there is anything in the new theory that they might be identified with or reduced to. If the posits of the new theory strike us as deeply and fundamentally different from those of the old theory, in the way that molecular motion seems deeply and fundamentally different from the ‘exquisitely elastic’ fluid posited by caloric theory, then it will be plausible to conclude that the theory change has been a radical one, and that an eliminativist conclusion is in order. But since there is no easy measure of how ‘deeply and fundamentally different’ a pair of posits are, the conclusion we reach is bound to be a judgment call.⁶

To argue that certain sorts of connectionist models support eliminativism about the propositional attitudes, we must make it plausible that these models are not ontologically conservative. Our strategy will be to contrast these connectionist models, models like those set out in Section 5, with ontologically conservative models like the one sketched at the end of Section 3, in an effort to underscore just how ontologically radical the connectionist models are. But here we are getting ahead of ourselves. Before trying to persuade you that connectionist models are ontologically radical, we need to take a look at the folk psychological theory that the connectionist models threaten to replace.

3. PROPOSITIONAL ATTITUDES AND COMMON SENSE PSYCHOLOGY

For present purposes we will assume that common sense psychology can plausibly be regarded as a theory, and that beliefs, desires and the rest of the propositional attitudes are plausibly viewed as posits of that theory. Though this is not an uncontroversial assumption, the case for it has been well argued by others.⁷ Once it is granted that common sense psychology is indeed a

theory, we expect it will be conceded by almost everyone that the theory is a likely candidate for replacement. In saying this, we do not intend to disparage folk psychology, or to beg any questions about the status of the entities it posits. Our point is simply that folk wisdom on matters psychological is not likely to tell us all there is to know. Common sense psychology, like other folk theories, is bound to be incomplete in many ways, and very likely to be inaccurate in more than a few. If this were not the case, there would be no need for a careful, quantitative, experimental science of psychology. With the possible exception of a few diehard Wittgensteinians, just about everyone is prepared to grant that there are many psychological facts and principles beyond those embedded in common sense. If this is right, then we have the first premise needed in an eliminativist argument aimed at beliefs, propositional memories and the rest of the propositional attitudes. The theory that posits the attitudes is indeed a prime candidate for replacement.

Though common sense psychology contains a wealth of lore about beliefs, memories, desires, hopes, fears and the other propositional attitudes, the crucial folk psychological tenets in forging the link between connectionism and eliminativism are the claims that propositional attitudes are *functionally discrete*, *semantically interpretable*, states that play a *causal role* in the production of other propositional attitudes, and ultimately in the production of behavior. Following the suggestion in Stich (1983), we'll call this cluster of claims *propositional modularity*.⁸ (The reader is cautioned not to confuse this notion of propositional modularity with the very different notion of modularity defended in Fodor (1983).)

The fact that common sense psychology takes beliefs and other propositional attitudes to have semantic properties deserves special emphasis. According to common sense:

- (i) when people see a dog nearby they typically come to believe that *there is a dog nearby*;
- (ii) when people believe that *the train will be late if there is snow in the mountains*, and come to believe that *there is snow in the mountains*, they will typically come to believe that *the train will be late*;
- (iii) when people who speak English say 'There is a cat in the yard,' they typically believe that *there is a cat in the yard*.

And so on, for indefinitely many further examples. Note that these generalizations of common sense psychology are couched in terms of the *semantic* properties of the attitudes. It is in virtue of being the belief that *p* that a given

belief has a given effect or cause. Thus common sense psychology treats the predicates expressing these semantic properties, predicates like ‘believes that the train is late’, as *projectable* predicates – the sort of predicates that are appropriately used in nomological or law-like generalizations.

There is a great deal of evidence that might be cited in support of the thesis that folk psychology is committed to the tenets of propositional modularity. Perhaps the most obvious way to bring out folk psychology’s commitment to the thesis that propositional attitudes are *functionally discrete* states is to note that it typically makes perfectly good sense to claim that a person has acquired (or lost) a single memory or belief. Thus, for example, on a given occasion it might plausibly be claimed that when Henry awoke from his nap he had completely forgotten that the car keys were hidden in the refrigerator, though he had forgotten nothing else. In saying that folk psychology views beliefs as the sorts of things that can be acquired or lost one at a time, we do not mean to be denying that having any particular belief may presuppose a substantial network of related beliefs. The belief that the car keys are in the refrigerator is not one that could be acquired by a primitive tribesman who knows nothing about cars, keys or refrigerators. But once the relevant background is in place, as we may suppose it is for us and for Henry, it seems that folk psychology is entirely comfortable with the possibility that a person may acquire (or lose) the belief that the car keys are in the refrigerator, while the remainder of his beliefs remain unchanged. Propositional modularity does not, of course, deny that acquiring one belief often leads to the acquisition of a cluster of related beliefs. When Henry is told that the keys are in the refrigerator, he may come to believe that they haven’t been left in the ignition, or in his jacket pocket. But then again he may not. Indeed, on the folk psychological conception of belief it is perfectly possible for a person to have a long standing belief that the keys are in the refrigerator, and to continue searching for them in the bedroom.⁹

To illustrate the way in which folk psychology takes propositional attitudes to be functionally discrete, *causally active* states let us sketch a pair of more elaborate examples.

(i) In common sense psychology, behavior is often explained by appeal to certain of the agent’s beliefs and desires. Thus, to explain why Alice went to her office, we might note that she wanted to send some e-mail messages (and, of course, she believed she could do so from her office). However, in some cases an agent will have several sets of beliefs and desires each of which *might* lead to the same behavior. Thus we may suppose that Alice also

wanted to talk to her research assistant, and that she believed he would be at the office. In such cases, common sense psychology assumes that Alice's going to her office might have been caused by either one of the belief/desire pairs, or by both, and that determining which of these options obtains is an empirical matter. So it is entirely possible that on *this* occasion Alice's desire to send some e-mail played no role in producing her behavior; it was the desire to talk with her research assistant that actually caused her to go to the office. However, had she not wanted to talk with her research assistant, she might have gone to the office anyhow, because the desire to send some e-mail, which was causally inert in her actual decision making, might then have become actively involved. Note that in this case common sense psychology is prepared to recognize a pair of quite distinct semantically characterized states, one of which may be causally active while the other is not.

(ii) Our second illustration is parallel to the first, but focuses on beliefs and inference, rather than desires and action. On the common sense view, it may sometimes happen that a person has a number of belief clusters, any one of which might lead him to infer some further belief. When he actually does draw the inference, folk psychology assumes that it is an empirical question what he inferred it from, and that this question typically has a determinate answer. Suppose, for example, that Inspector Clouseau believes that the butler said he spent the evening at the village hotel, and that he said he arrived back on the morning train. Suppose Clouseau also believes that the village hotel is closed for the season, and that the morning train has been taken out of service. Given these beliefs, along with some widely shared background beliefs, Clouseau might well infer that the butler is lying. If he does, folk psychology presumes that the inference might be based either on his beliefs about the hotel, or on his beliefs about the train, or both. It is entirely possible, from the perspective of common sense psychology, that although Clouseau has long known that the hotel is closed for the season, this belief played no role in his inference on this particular occasion. Once again we see common sense psychology invoking a pair of distinct propositional attitudes, one of which is causally active on a particular occasion while the other is causally inert.

In the psychological literature there is no shortage of models for human belief or memory which follows the lead of common sense psychology in supposing that propositional modularity is true. Indeed, prior to the emergence of connectionism, just about all psychological models of propositional

memory, save for those urged by behaviorists, were comfortably compatible with propositional modularity. Typically, these models view a subject's store of beliefs or memories as an interconnected collection of functionally discrete, semantically interpretable states which interact in systematic ways. Some of these models represent individual beliefs as sentence-like structures – strings of symbols which can be individually activated by transferring them from long term memory to the more limited memory of a central processing unit. Other models represent beliefs as a network of labeled nodes and labeled links through which patterns of activation may spread. Still other models represent beliefs as sets of production rules.¹⁰ In all three sorts of models, it is generally the case that for any given cognitive episode, like performing a particular inference or answering a question, some of the memory states will be actively involved, and others will be dormant.

SEMANTIC NETWORK

PROPOSITIONS

1. Dogs have fur.
2. Dogs have paws.
3. Cats have fur.
4. Cats have paws.

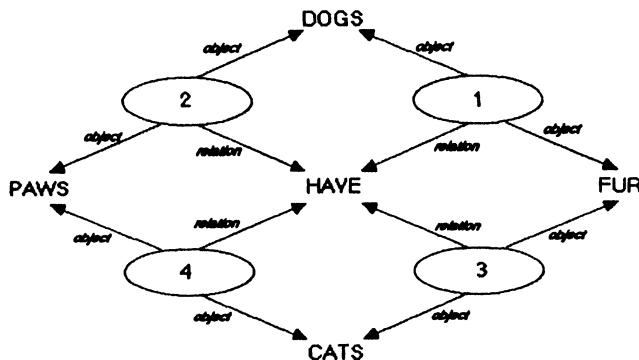


Fig. 1.

SEMANTIC NETWORK

PROPOSITIONS

1. Dogs have fur.
2. Dogs have paws.
3. Cats have fur.

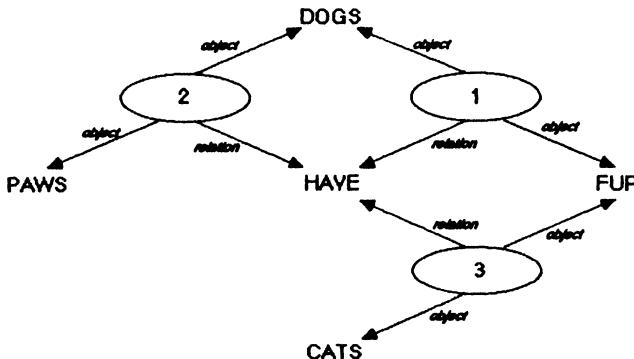


Fig. 2.

In Figure 1 we have displayed a fragment of a ‘semantic network’ representation of memory, in the style of Collins and Quillian (1972). In this model, each distinct proposition in memory is represented by an oval node along with its labeled links to various concepts. By adding assumptions about the way in which questions or other sorts of memory probes lead to activation spreading through the network, the model enables us to make predictions about speed and accuracy in various experimental studies of memory. For our purposes there are three facts about this model that are of particular importance. First, since each proposition is encoded in a functionally discrete way, it is a straightforward matter to add or subtract a *single* proposition from memory while leaving the rest of the network unchanged. Thus, for example, Figure 2 depicts the result of removing one proposition from the network in Figure 1. Second, the model treats predicates expressing the semantic properties of beliefs or memories as *projectable*.¹¹ They are treated as the sorts of predicates that pick out scientifically genuine *kinds*, rather than mere accidental conglomerates, and thus are suitable for inclusion in the statement

of lawlike regularities. To see this, we need only consider the way in which such models are tested against empirical data about memory acquisition and forgetting. Typically, it will be assumed that if a subject is told (for example) that the policeman arrested the hippie, then the subject will (with a certain probability) remember that *the policeman arrested the hippie*.¹² And this assumption is taken to express a nomological generalization – it captures something lawlike about the way in which the cognitive system works. So while the class of people who *remember that the policeman arrested the hippie* may differ psychologically in all sorts of ways, the theory treats them as a psychologically natural kind. Third, in any given memory search or inference task exploiting a semantic network model, it makes sense to ask which propositions were activated and which were not. Thus, a search in the network of Figure 1 might terminate without ever activating the proposition that cats have paws.

4. A FAMILY OF CONNECTIONIST HYPOTHESES

Our theme, in the previous section, was that common sense psychology is committed to propositional modularity, and that many models of memory proposed in the cognitive psychology literature are comfortably compatible with this assumption. In the present section we want to describe a class of connectionist models which, we will argue, are *not* readily compatible with propositional modularity. The connectionist models we have in mind share three properties:

- (i) their encoding of information in the connection weights and in the biases on units is *widely distributed*, rather than being *localist*;
- (ii) individual hidden units in the network have no comfortable symbolic interpretation; they are *subsymbolic*, to use a term suggested by Paul Smolensky;
- (iii) the models are intended as *cognitive models*, not merely as *implementations* of cognitive models.

A bit later in this section we will elaborate further on each of these three features, and in the next section we will describe a simple example of a connectionist model that meets our three criteria. However, we are not under any illusion that what we say will be sufficient to give a sharp edged characterization of the class of connectionist models we have in mind. Nor is

such a sharp edged characterization essential for our argument. It will suffice if we can convince you that there is a significant class of connectionist models which are incompatible with the propositional modularity of folk psychology.

Before saying more about the three features on our list, we would do well to give a more general characterization of the sort of models we are calling 'connectionist', and introduce some of the jargon that comes with the territory. To this end, let us quote at some length from Paul Smolensky's lucid overview.

Connectionist models are large networks of simple, parallel computing elements, each of which carries a numerical *activation value* which it computes from neighboring elements in the network, using some simple numerical formula. The network elements or *units* influence each other's values through connections that carry a numerical strength or *weight*...

In a typical ... model, input to the system is provided by imposing activation values on the *input units* of the network; these numerical values represent some encoding or *representation* of the input. The activation on the input units propagates along the connections until some set of activation values emerges on the *output units*; these activation values encode the output the system has computed from the input. In between the input and output units there may be other units, often called *hidden units*, that participate in representing neither the input nor the output.

The computation performed by the network in transforming the input pattern of activity to the output pattern depends on the set of connection strengths; *these weights are usually regarded as encoding the system's knowledge*.¹³ In this sense, the connection strengths play the role of the program in a conventional computer. Much of the allure of the connectionist approach is that many connectionist networks *program themselves*, that is, they have autonomous procedures for tuning their weights to eventually perform some specific computation. Such *learning procedures* often depend on training in which the network is presented with sample input/output pairs from the function it is supposed to compute. In learning networks with hidden units, the network itself 'decides' what computations the hidden units will perform; because these units represent neither inputs nor outputs, they are never 'told' what their values should be, even during training...¹⁴

One point must be added to Smolensky's portrait. In many connectionist models the hidden units and the output units are assigned a numerical 'bias' which is added into the calculation determining the unit's activation level. The learning procedures for such networks typically set both the connection strengths and the biases. Thus in these networks the system's knowledge is usually regarded as encoded in *both* the connection strengths and the biases.

So much for a general overview. Let us now try to explain the three features that characterize those connectionist models we take to be incompatible with propositional modularity.

(i) In many non-connectionist cognitive models, like the one illustrated at the end of Section 3, it is an easy matter to locate a functionally distinct part of the model encoding each proposition or state of affairs represented in the system. Indeed, according to Fodor and Pylyshyn, “conventional [computations] architecture requires that there be distinct symbolic expressions for each state of affairs that it can represent.”¹⁵ In some connectionist models an analogous sort of functional localization is possible, not only for the input and output units but for the hidden units as well. Thus, for example, in certain connectionist models, various individual units or small clusters of units are themselves intended to represent specific properties or features of the environment. When the connection strength from one such unit to another is strongly positive, this might be construed as the system’s representation of the proposition that if the first feature is present, so too is the second. However, in many connectionist networks it is not possible to localize propositional representation beyond the input layer. That is, there are no particular features or states of the system which lend themselves to a straightforward semantic evaluation. This can sometimes be a real inconvenience to the connectionist model builder when the system as a whole fails to achieve its goal because it has not represented the world the way it should. When this happens, as Smolensky notes,

[I]t is not necessarily possible to localize a failure of veridical representation. Any particular state is part of a large causal system of states, and failures of the system to meet goal conditions cannot in general be localized to any particular state or state component.¹⁶

It is connectionist networks of this sort, in which it is not possible to isolate the representation of particular propositions or states of affairs within the nodes, connection strengths and biases, that we have in mind when we talk about the encoding of information in the biases, weights and hidden nodes being *widely distributed* rather than *localist*.

(ii) As we’ve just noted, there are some connectionist models in which some or all of the units are intended to represent specific properties or features of the system’s environment. These units may be viewed as the model’s symbols for the properties or features in question. However, in models where the weights and biases have been tuned by learning algorithms it is often not the case that any single unit or any small collection of units will end up representing a specific feature of the environment in any straightforward way. As we shall see in the next section, it is often plausible to view

such networks as collectively or holistically encoding a set of propositions, although none of the hidden units, weights or biases are comfortably viewed as *symbols*. When this is the case we will call the strategy of representation invoked in the model *subsymbolic*. Typically (perhaps always?) networks exploiting subsymbolic strategies of representation will encode information in a widely distributed way.

(iii) The third item on our list is not a feature of connectionist models themselves, but rather a point about how the models are to be interpreted. In making this point we must presuppose a notion of theoretical or explanatory level which, despite much discussion in the recent literature, is far from being a paradigm of clarity.¹⁷ Perhaps the clearest way to introduce the notion of explanatory level is against the background of the familiar functionalist thesis that psychological theories are analogous to programs which can be implemented on a variety of very different sorts of computers.¹⁸ If one accepts this analogy, then it makes sense to ask whether a particular connectionist model is intended as a model at the psychological level or at the level of underlying neural implementation. Because of their obvious, though in many ways very partial, similarity to real neural architectures, it is tempting to view connectionist models as models of the implementation of psychological processes. And some connectionist model builders endorse this view quite explicitly. So viewed, however, connectionist models are not *psychological* or *cognitive* models at all, any more than a story of how cognitive processes are implemented at the quantum mechanical level is a psychological story. A very different view that connectionist model builders can and often do take is that their models are at the psychological level, not at the level of implementation. So construed, the models are in competition with other psychological models of the same phenomena. Thus a connectionist model of word recognition would be an alternative to – and not simply a possible implementation of – a non-connectionist model of word recognition; a connectionist theory of memory would be a competitor to a semantic network theory, and so on. Connectionists who hold this view of their theories often illustrate the point by drawing analogies with other sciences. Smolensky, for example, suggests that connectionist models stand to traditional cognitive models (like semantic networks) in much the same way that quantum mechanics stands to classical mechanics. In each case the newer theory is deeper, more general and more accurate over a broader range of phenomena. But in each case the new theory and the old are competing at the same explanatory level. If one is right, the other must be wrong.

In light of our concerns in this paper, there is one respect in which the analogy between connectionist models and quantum mechanics may be thought to beg an important question. For while quantum mechanics is conceded to be a *better* theory than classical mechanics, a plausible case could be made that the shift from classical to quantum mechanics was an ontologically *conservative* theory change. In any event, it is not clear that the change was ontologically *radical*. If our central thesis in this paper is correct, then the relation between connectionist models and more traditional cognitive models is more like the relation between the caloric theory of heat and the kinetic theory. The caloric and kinetic theories are at the same explanatory level, though the shift from one to the other was pretty clearly ontologically radical. In order to make the case that the caloric analogy is the more appropriate one, it will be useful to describe a concrete, though very simple, connectionist model of memory that meets the three criteria we have been trying to explicate.

5. A CONNECTIONIST MODEL OF MEMORY

Our goal in constructing the model was to produce a connectionist network that would do at least some of the tasks done by more traditional cognitive models of memory, and that would perspicuously exhibit the sort of distributed, sub-symbolic encoding described in the previous section. We began by constructing a network, we'll call it Network A, that would judge the truth or falsehood of the sixteen propositions displayed above the line in Figure 3. The network was a typical three tiered feed-forward network consisting of 16 input units, four hidden units and one output unit, as shown in Figure 4. The input coding of each proposition is shown in the center column in Figure 3. Outputs close to 1 were interpreted as 'true' and outputs close to zero were interpreted as 'false'. Back propagation, a familiar connectionist learning algorithm, was used to 'train up' the network thereby setting the connection weights and biases. Training was terminated when the network consistently gave an output higher than 0.9 for each true proposition and lower than 0.1 for each false proposition. Figure 5 shows the connection weights between the input units and the leftmost hidden unit in the trained up network, along with the bias on that unit. Figure 6 indicates the connection weights and biases further upstream. Figure 7 shows the way in which the network computes its response to the proposition *Dogs have fur* when that proposition is encoded in the input units.

Proposition	Input	Output
1 Dogs have fur.	11000011 00001111	1 true
2 Dogs have paws.	11000011 00110011	1 true
3 Dogs have fleas.	11000011 00111111	1 true
4 Dogs have legs.	11000011 00111100	1 true
5 Cats have fur.	11001100 00001111	1 true
6 Cats have paws.	11001100 00110011	1 true
7 Cats have fleas.	11001100 00111111	1 true
8 Fish have scales.	11110000 00110000	1 true
9 Fish have fins.	11110000 00001100	1 true
10 Fish have gills.	11110000 00000011	1 true
11 Cats have gills.	11001100 00000011	0 false
12 Fish have legs.	11110000 00111100	0 false
13 Fish have fleas.	11110000 00111111	0 false
14 Dogs have scales.	11000011 00110000	0 false
15 Dogs have fins.	11000011 00001100	0 false
16 Cats have fins.	11001100 00001100	0 false
<hr/>		
Added Proposition		
17 Fish have eggs.	11110000 11001000	1 true

Fig. 3.

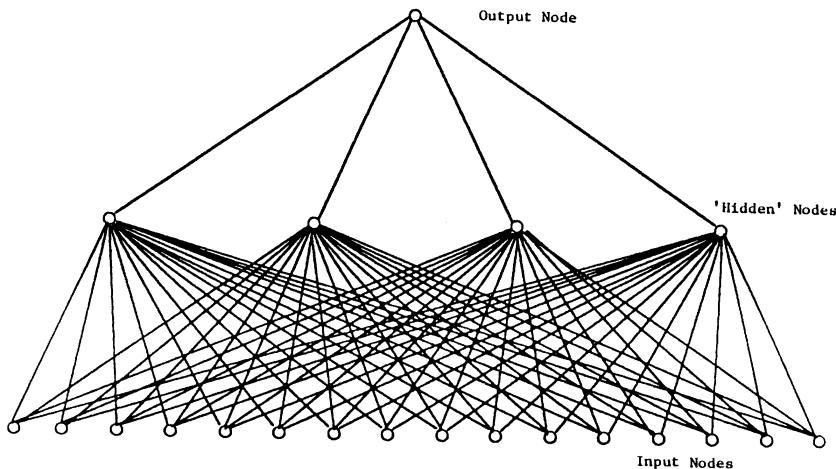
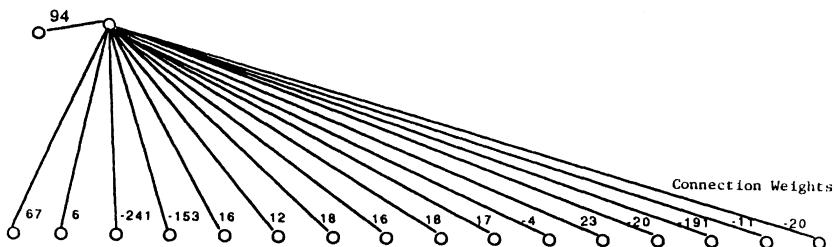
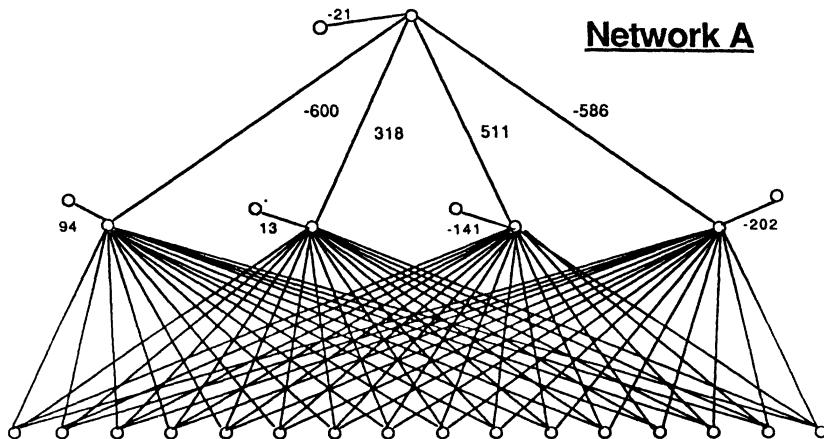


Fig. 4.

Network A

Input weights and bias to first hidden node
in network with 16 propositions.

Fig. 5.



Weights and biases in network
with 16 propositions.

Fig. 6.

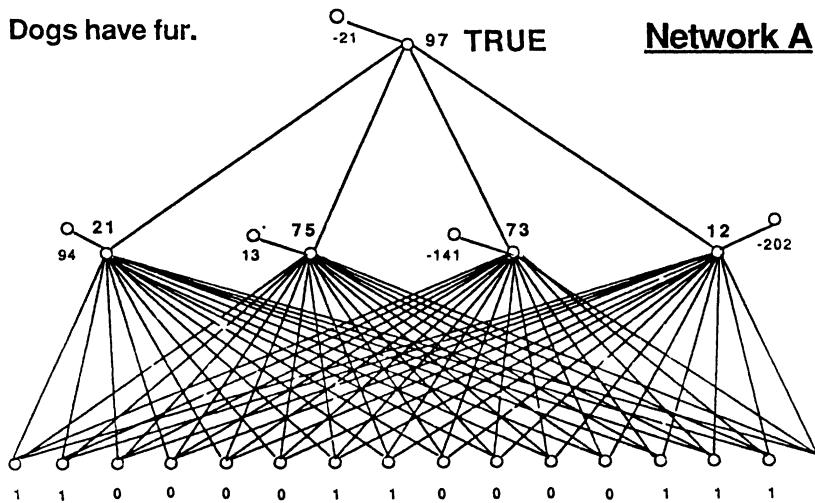


Fig. 7.

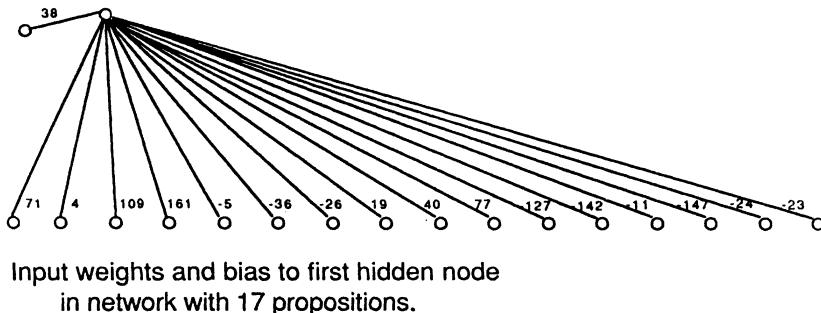
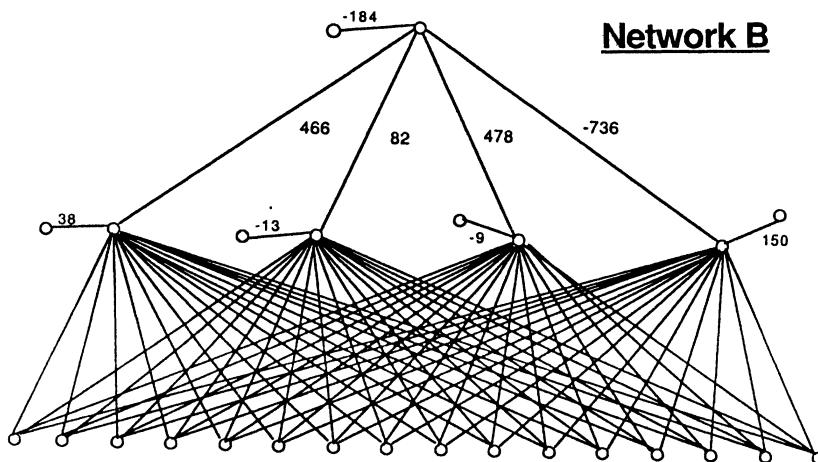
Network B

Fig. 8.



Weights and biases in network
with 17 propositions.

Fig. 9.

There is a clear sense in which the trained up Network A may be said to have stored information about the truth or falsity of propositions (1)–(16), since when any one of these propositions is presented to the network it correctly judges whether the proposition is true or false. In this respect it is similar to various semantic network models which can be constructed to perform much the same task. However, there is a striking difference between Network A and a semantic network model like the one depicted in Figure 1. For, as we noted earlier, in the semantic network there is a functionally distinct sub-part associated with each proposition, and thus it makes perfectly good sense to ask, for any probe of the network, whether or not the representation of a specific proposition played a causal role. In the connectionist network, by contrast, there is no distinct state or part of the network that serves to represent any particular proposition. The information encoded in Network A is stored holistically and distributed throughout the network. Whenever information is extracted from Network A, by giving it an input string and seeing whether it computes a high or a low value for the output

unit, *many* connection strengths, *many* biases and *many* hidden units play a role in the computation. And any particular weight or unit or bias will help to encode information about *many* different propositions. It simply makes no sense to ask whether or not the representation of a particular proposition plays a causal role in the network's computation. It is in just this respect that our connectionist model of memory seems radically incongruent with the propositional modularity of common sense psychology. For, as we saw in Section 3, common sense psychology seems to presuppose that there is generally some answer to the question of whether a particular belief or memory played a causal role in a specific cognitive episode. But if belief and memory are subserved by a connectionist network like ours, such questions seem to have no clear meaning.

The incompatibility between propositional modularity and connectionist models like ours can be made even more vivid by contrasting Network A with a second network, we'll call it Network B, depicted in Figures 8 and 9. Network B was trained up just as the first one was, except that one additional proposition was added to the training set (coded as indicated below the line in Figure 3). Thus Network B encodes all the same propositions as Network A plus one more. In semantic network models, and other traditional cognitive models, it would be an easy matter to say which states or features of the system encode the added proposition, and it would be a simple task to determine whether or not the representation of the added proposition played a role in a particular episode modeled by the system. But plainly in the connectionist network those questions are quite senseless. The point is not that there are no differences between the two networks. Quite the opposite is the case; the differences are many and widespread. But these differences do not correlate in any systematic way with the functionally discrete, semantically interpretable states posited by folk psychology and by more traditional cognitive models. Since information is encoded in a highly distributed manner, with each connection weight and bias embodying information salient to many propositions, and with information regarding any given proposition scattered throughout the network, the system lacks functionally distinct, identifiable sub-structures that are semantically interpretable as representations of individual propositions.

The contrast between Network A and Network B enables us to make our point about the incompatibility between common sense psychology and these sorts of connectionist models in a rather different way. We noted in Section 3 that common sense psychology treats predicates expressing the semantic properties of propositional attitudes as projectable. Thus 'believes that dogs

have fur' or 'remembers that dogs have fur' will be projectable predicates in common sense psychology. Now both Network A and Network B might serve as models for a cognitive agent who believes that dogs have fur; both networks store or represent the information that dogs have fur. Nor are these the only two. If we were to train up a network on the 17 propositions in Figure 3 plus a few (or minus a few) we would get yet another system which is as different from Network A and B as these two are from each other. The moral here is that though there are *indefinitely* many connectionist networks that represent the information that dogs have fur just as well as Network A does, these networks have no projectable features in common that are describable in the language of connectionist theory. From the point of view of the connectionist model builder, the class of networks that might model a cognitive agent who believes that dogs have fur is not a genuine kind at all, but simply a chaotically disjunctive set. Common sense psychology treats the class of people who believe that dogs have fur as a psychologically natural kind; connectionist psychology does not.¹⁹

6. OBJECTIONS AND REPLIES

The argument we've set out in the previous five sections has encountered no shortage of objections. In this section we will try to reconstruct the most interesting of these, and indicate how we would reply.

Objection (i): Models like A and B are not serious models for human belief or propositional memory.

Of course, the models we've constructed are tiny toys that were built to illustrate the features set out in Section 4 in a perspicuous way. They were never intended to model any substantial part of human propositional memory. But various reasons have been offered for doubting that *anything like* these models could ever be taken seriously as psychological models of propositional memory. Some critics have claimed that the models simply will not scale up – that while teaching a network to recognize fifteen or twenty propositions may be easy enough, it is just not going to be possible to train up a network that can recognize a few thousand propositions, still less a few hundred thousand.²⁰ Others have objected that while more traditional models of memory, including those based on sentence-like storage, those using semantic networks, and those based on production systems, all provide some strategy for *inference* or *generalization* which enables the system to answer questions about propositions it was not explicitly taught, models like those

we have constructed are incapable of inference and generalization. It has also been urged that these models fail as accounts of human memory because they provide no obvious way to account for the fact that suitably prepared humans can easily acquire propositional information one proposition at a time. Under ordinary circumstances, we can just *tell* Henry that the car keys are in the refrigerator, and he can readily record this fact in memory. He doesn't need anything like the sort of massive retraining that would be required to teach one of our connectionist networks a new proposition.

Reply: If this were a paper aimed at defending connectionist models of propositional memory, we would have to take on each of these putative shortcomings in some detail. And in each instance there is at least something to be said on the connectionist side. Thus, for example, it just is not true that networks like A and B don't generalize beyond the propositions on which they've been trained. In Network A, for example, the training set included:

Dogs have fur	Cats have fur
Dogs have paws	Cats have paws
Dogs have fleas	Cats have fleas.

It also included

Dogs have legs

but not

Cats have legs.

When the network was given an encoding of this last proposition, however, it generalized correctly and responded affirmatively. Similarly, the network responded negatively to an encoding of

Cats have scales

though it had not previously been exposed to this proposition.

However, it is important to see that this sort of point-by-point response to the charge that networks like ours are inadequate models for propositional memory is not really required, given the thesis we are defending in this paper. For what we are trying to establish is a *conditional* thesis: *if* connectionist models of memory of the sort we describe in Section 4 are right, *then* propositional attitude psychology is in serious trouble. Since conditionals with false antecedents are true, we win by default if it turns out that the antecedent of our conditional is false.

Objection (ii): Our models do not really violate the principle of propositional modularity, since the propositions the system has learned are coded in functionally discrete ways, though this may not be obvious.

We've heard this objection elaborated along three quite different lines. The first line – let's call it Objection (iia) – notes that functionally discrete coding may often be *very* hard to notice, and can not be expected to be visible on casual inspection. Consider, for example, the way in which sentences are stored in the memory of a typical von Neumann architecture computer – for concreteness we might suppose that the sentences are part of an English text and are being stored while the computer is running a word processing program. Parts of sentences may be stored at physically scattered memory addresses linked together in complex ways, and given an account of the contents of all relevant memory addresses one would be hard put to say where a particular sentence is stored. But nonetheless each sentence is stored in a *functionally discrete* way. Thus if one knew enough about the system it would be possible to erase any particular sentence it is storing by tampering with the contents of the appropriate memory addresses, while leaving the rest of the sentences the system is storing untouched. Similarly, it has been urged, connectionist networks may in fact encode propositions in functionally discrete ways, though this may not be evident from a casual inspection of the trained up network's biases and connection strengths.

Reply (iia): It is a bit difficult to come to grips with this objection, since what the critic is proposing is that in models like those we have constructed there *might* be some covert functionally discrete system of propositional encoding that has yet to be discovered. In response to this we must concede that indeed there might. We certainly have no argument that even comes close to demonstrating that the discovery of such a covert functionally discrete encoding is impossible. Moreover, we concede that if such a covert system were discovered, then our argument would be seriously undermined. However, we're inclined to think that the burden of argument is on the critic to show that such a system is not merely possible but *likely*; in the absence of any serious reason to think that networks like ours do encode propositions in functionally discrete ways, the mere logical possibility that they might is hardly a serious threat.

The second version of Objection (ii) – we'll call it Objection (iib) – makes a specific proposal about the way in which networks like A and B might be discretely, though covertly, encoding propositions. The encoding, it is urged,

is to be found in the pattern of activation of the hidden nodes, when a given proposition is presented to the network. Since there are four hidden nodes in our networks, the activation pattern on presentation of any given input may be represented as an ordered 4-tuple. Thus, for example, when network A is presented with the encoded proposition *Dogs have fur*, the relevant 4-tuple would be (21, 75, 73, 12), as shown in Figure 7. Equivalently, we may think of each activation pattern as a point in a four dimensional hyperspace. Since each proposition corresponds to a unique point in the hyperspace, that point may be viewed as the encoding of the proposition. Moreover, that point represents a functionally discrete state of the system.²¹

Reply (iib): What is being proposed is that the pattern of activation of the system on presentation of an encoding of the proposition p be identified with the belief that p . But this proposal is singularly implausible. Perhaps the best way to see this is to note that in common sense psychology beliefs and propositional memories are typically of substantial duration; and they are the sorts of things that cognitive agents generally have lots of even when they are not using them. Consider an example. Are kangaroos marsupials? Surely you've believed for years that they are, though in all likelihood this is the first time today that your belief has been activated or used.²² An activation pattern, however, is not an enduring state of a network; indeed, it is not a state of the network at all except when the network has had the relevant proposition as input. Moreover, there is an enormous number of other beliefs that you've had for years. But it makes no sense to suppose that a network could have many activation patterns continuously over a long period of time. At any given time a network exhibits at most one pattern of activation. So activation patterns are just not the sorts of things that can plausibly be identified with beliefs or their representations.

Objection (iic): At this juncture, a number of critics have suggested that long standing beliefs might be identified not with activation patterns, which are transient states of networks, but rather with *dispositions to produce activation patterns*. Thus, in network A, the belief that dogs have fur would not be identified with a location in activation hyperspace but with the network's *disposition* to end up at that location when the proposition is presented. This *dispositional state* is an enduring state of the system; it is a state the network can be in no matter what its current state of activation may be, just as a sugar cube may have a disposition to dissolve in water even when there is no water nearby.²³ Some have gone on to suggest that the familiar philosophical distinction between dispositional and occurrent beliefs might be captured, in connectionist models, as the distinction between

dispositions to produce activation patterns and activation patterns themselves.

Reply (iic): Our reply to this suggestion is that while dispositions to produce activation patterns are indeed *enduring* states of the system, they are not the right sort of enduring states – they are not the discrete, independently causally active states that folk psychology requires. Recall that on the folk psychological conception of belief and inference, there will often be a variety of quite different underlying causal patterns that may lead to the acquisition and avowal of a given belief. When Clouseau says that the butler did it, he may have just inferred this with the help of his long standing belief that the train is out of service. Or he may have inferred it by using his belief that the hotel is closed. Or both long standing beliefs may have played a role in the inference. Moreover, it is also possible that Clouseau drew this inference some time ago, and is now reporting a relatively long standing belief. But it is hard to see how anything like these distinctions can be captured by the dispositional account in question. In reacting to a given input, say p , a network takes on a specific activation value. It may also have dispositions to take on other activation values on other inputs, say q and r . But there is no obvious way to interpret the claim that these further dispositions play a causal role in the network's reaction to p – or, for that matter, that they do not play a role. Nor can we make any sense of the idea that on one occasion the encoding of q (say, the proposition that the train is out of service) played a role while the encoding of r (say, the proposition that the hotel is closed) did not, and on another occasion, things went the other way around. The propositional modularity presupposed by common sense psychology requires that belief tokens be functionally discrete states capable of causally interacting with one another in some cognitive episodes and of remaining causally inert in other cognitive episodes. However, in a distributed connectionist system like Network A, the dispositional state which produces one activation pattern is functionally inseparable from the dispositional state which produces another. Thus it is impossible to isolate some propositions as causally active in certain cognitive episodes, while others are not. We conclude that reaction pattern dispositions won't do as belief tokens. Nor, so far as we can see, are there any other states of networks like A and B that will fill the bill.

7. CONCLUSION

The thesis we have been defending in this paper is that connectionist models of a certain sort are incompatible with the propositional modularity em-

bedded in common sense psychology. The connectionist models in question are those which are offered as models at the *cognitive* level, and in which the encoding of information is widely distributed and subsymbolic. In such models, we have argued, there are no *discrete, semantically interpretable* states that play a *causal role* in some cognitive episodes but not others. Thus there is, in these models, nothing with which the propositional attitudes of common sense psychology can plausibly be identified. If these models turn out to offer the best accounts of human belief and memory, we will be confronting an *ontologically radical* theory change – the sort of theory change that will sustain the conclusion that propositional attitudes, like caloric and phlogiston, do not exist.

NOTES

¹ Thanks are due to Ned Block, Paul Churchland, Gary Cottrell, Adrian Cussins, Jerry Fodor, John Heil, Frank Jackson, David Kirsh, Patricia Kitcher and Philip Kitcher for useful feedback on earlier versions of this paper. Talks based on the paper have been presented at the UCSD Cognitive Science Seminar and at conferences sponsored by the Howard Huges Medical Foundation and the University of North Carolina at Greensboro. Comments and questions from these audiences have proved helpful in many ways.

² See, for example, Churchland (1981) and (1986), where explicitly eliminativist conclusions are drawn on the basis of speculations about the success of cognitive models similar to those we shall discuss.

³ Fodor, J. and Pylyshyn, Z. (1988).

⁴ We are aware that certain philosophers and historians of science have actually entertained ideas similar to the suggestion that the planets spoken of by pre-Copernican astronomers do not exist. See, for example, Kuhn (1970), Ch. 10, and Feyerabend (1981), Ch. 4. However, we take this suggestion to be singularly implausible. Eliminativist arguments can't be that easy. Just what has gone wrong with the accounts of meaning and reference that lead to such claims is less clear. For further discussion on these matters see Kuhn (1983), and Kitcher (1978) and (1983).

⁵ For some detailed discussion of scientific reduction, see Nagel (1961); Schaffner (1968); Hooker (1981); and Kitcher (1984). The genetics case is not without controversy. See Kitcher (1982) and (1984).

⁶ It's worth noting that judgments on this matter can differ quite substantially. At one end of the spectrum are writers like Feyerabend (1981), and perhaps Kuhn (1962), for whom relatively small differences in theory are enough to justify the suspicion that there has been an ontologically radical change. Toward the other end are writers like Lycan, who writes:

I am at pains to advocate a very liberal view... I am entirely willing to give up fairly large chunks of our commonsensical or platitudinous theory of belief or of desire (or of almost anything else) and decide that we were just wrong about a lot of things,

without drawing the inference that we are no longer talking about belief or desire... I think the ordinary word 'belief' (*qua* theoretical term of folk psychology) points dimly toward a natural kind that we have not fully grasped and that only mature psychology will reveal. I expect that 'belief' will turn out to refer to some kind of information bearing inner state of a sentient being ..., but the kind of state it refers to may have only a few of the properties usually attributed to beliefs by common sense. (Lycan (1988), pp. 31-2.)

On our view, both extreme positions are implausible. As we noted earlier, the Copernican revolution did not show that the planets studied by Ptolemy do not exist. But Lavoisier's chemical revolution *did* show that phlogiston does not exist. Yet on Lycan's "very liberal view" it is hard to see why we should not conclude that phlogiston really does exist after all – it's really oxygen, and prior to Lavoisier 'we were just very wrong about a lot of things'.

⁷ For an early and influential statement of the view that common sense psychology is a theory, see Sellars (1956). More recently the view has been defended by Churchland (1970) and (1979), Chs. 1 and 4; and by Fodor (1988), Ch. 1. For the opposite view, see Wilkes (1978); Madell (1986); Sharpe (1987).

⁸ See Stich (1983), pp. 237 ff.

⁹ Cherniak (1986), Ch. 3, notes that this sort of absent mindedness is commonplace in literature and in ordinary life, and sometimes leads to disastrous consequences.

¹⁰ For sentential models, see John McCarthy (1968), (1980), and (1986); and Kintsch (1974). For semantic networks, see Quillian (1969); Collins and Quillian (1972); Rumelhart, Lindsay and Norman (1972); Anderson and Bower (1973); and Anderson (1976) and (1980), Ch. 4. For production systems, see Newell and Simon (1972); Newell (1973); Anderson (1983); and Holland *et al.* (1986).

¹¹ For the classic discussion of the distinction between projectable and non-projectable predicates, see Goodman (1965).

¹² See, for example, Anderson and Bower (1973).

¹³ Emphasis added.

¹⁴ Smolensky (1988), p. 1.

¹⁵ Fodor and Pylyshyn (1988), p. 57.

¹⁶ Smolensky (1988), p. 15.

¹⁷ Broadbent (1985); Rumelhart and McClelland (1985); Rumelhart and McClelland (1986), Ch. 4; Smolensky (1988); Fodor and Pylyshyn (1988).

¹⁸ The notion of program being invoked here is itself open to a pair of quite different interpretations. For the right reading, see Ramsey (1989).

¹⁹ This way of making the point about the incompatibility between connectionist models and common sense psychology was suggested to us by Jerry Fodor.

²⁰ This point has been urged by Daniel Dennett, among others.

²¹ Quite a number of people have suggested this move, including Gary Cottrell and Adrian Cussins.

²² As Lycan notes, on the common sense notion of belief, people have lots of them "even when they are asleep." (Lycan (1988), p. 57.)

²³ Something like this objection was suggested to us by Ned Block and by Frank Jackson.

REFERENCES

- Anderson, J. and Bower, G.: 1973, *Human Associative Memory*, Washington, D.C., Winston.
- Anderson, J.: 1976, *Language, Memory and Thought*, Hillsdale, N.J., Lawrence Erlbaum Associates.
- Anderson, J.: 1980, *Cognitive Psychology and Its Implications* San Francisco, W.H. Freeman & Co.
- Anderson, J.: 1983, *The Architecture of Cognition*, Cambridge, MA, Harvard University Press.
- Broadbent, D.: 1985, 'A Question of Levels: Comments on McClelland and Rumelhart', *Journal of Experimental Psychology: General*, 114.
- Cherniak, C.: 1986, *Minimal Rationality*, Cambridge, Mass., Bradford Books/MIT Press.
- Churchland, P.: 1970, 'The Logical Character of Action Explanations', *Philosophical Review*, 79.
- Churchland, P.: 1979, *Scientific Realism and the Plasticity of Mind*, Cambridge, Cambridge University Press.
- Churchland, P.: 1981, 'Eliminative Materialism and Propositional Attitudes', *Journal of Philosophy*, 78, 2.
- Churchland, P.: 1986, 'Some Reductive Strategies in Cognitive Neurobiology', *Mind*, 95.
- Collins, A. and Quillian, M.: 1972, 'Experiments on Semantic Memory and Language Comprehension', in L. Gregg, (ed.), *Cognition in Learning and Memory*, New York, Wiley.
- Feyerabend, P.: 1981, *Realism, Rationalism and Scientific Method: Philosophical Papers Vol. 1*, Cambridge, Cambridge University Press.
- Fodor, J. and Pylyshyn, Z.: 1988, 'Connectionism and Cognitive Architecture: A Critical Analysis', *Cognition*, 28.
- Fodor, J.: 1987, *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*, Cambridge, MA, Bradford Books/MIT Press.
- Goodman, N.: 1965, *Fact Fiction and Forecast*, Indianapolis, Bobbs-Merrill.
- Holland, J., Holyoak, K., Nisbett, R. and Thagard, P.: 1986, *Induction, Processes of Inference, Learning and Discovery*, Cambridge, MA, Bradford Books/MIT Press.
- Hooker, C.: 1981, 'Towards a General Theory of Reduction', Parts I, II & III, *Dialogue*, 20.
- Kintsch, W.: 1974, *The Representation of Meaning in Memory*, Hillsdale, N.J., Lawrence Erlbaum Associates.
- Kitcher, P.: 1978, 'Theories, Theorists and Theoretical Change', *Philosophical Review*, 87.
- Kitcher, P.: 1982, 'Genes', *British Journal for the Philosophy of Science*, 33.
- Kitcher, P.: 1983, 'Implications of Incommensurability', *PSA 1982* (Proceedings of the 1982 Biennial Meeting of the Philosophy of Science Association), Vol. 2, ed. by P. Asquith and T. Nickles, East Lansing, Philosophy of Science Association.
- Kitcher, P.: 1984, '1953 and All That: A Tale of Two Sciences', *Philosophical Review*, 93.
- Kuhn, T.: 1962, *The Structure of Scientific Revolutions*, Chicago, University of Chicago Press, 2nd Edition (1970).
- Kuhn, T.: 1983, 'Commensurability, Comparability, Communicability', *PSA 1982* (Proceedings of the 1982 Biennial Meeting of the Philosophy of Science Association)

- tion), Vol. 2, ed. by P. Asquith and T. Nickles, East Lansing, Philosophy of Science Association.
- Lycan, W.: 1988, *Judgment and Justification*, Cambridge, Cambridge University Press.
- Madell, G.: 1986, 'Neurophilosophy: A Principled Skeptic's Response', *Inquiry*, 29.
- McCarthy, J.: 1968, 'Programs With Common Sense', in M. Minsky, ed., *Semantic Information Processing*, Cambridge, MA, MIT Press.
- McCarthy, J.: 1980, 'Circumscription: A Form of Non-Monotonic Reasoning', *Artificial Intelligence*, 13.
- McCarthy, J.: 1986, 'Applications of Circumscription to Formalizing Common-Sense Knowledge', 28.
- Nagel, E.: 1961, *The Structure of Science*, New York, Harcourt, Brace & World.
- Newell, A. and Simon, H.: 1972, *Human Problem Solving*, Englewood Cliffs, N.J., Prentice Hall.
- Newell, A.: 1973, 'Production Systems: Models of Control Structures', in W. Chase (ed.), *Visual Information Processing*, New York, Academic Press.
- Quillian, M.: 1966, *Semantic Memory*, Cambridge, MA, Bolt, Branak & Newman.
- Ramsey, W.: 1989, 'Parallelism and Functionalism', *Cognitive Science*, 13.
- Rumelhart, D. and McClelland, J.: 1985, 'Level's Indeed! A Response to Broadbent', *Journal of Experimental Psychology: General*, 114.
- Rumelhart, D., Lindsay, P. and Norman, D.: 1972, 'A Process Model for Long Term Memory', in E. Tulving and W. Donaldson (eds.), *Organization of Memory*, New York, Academic Press.
- Rumelhart, D., McClelland, J. and the PDP Research Group (1986). *Parallel Distributed Processing*, Volumes I and II, Cambridge, MA, Bradford Books/MIT Press.
- Sellars, W.: 1956, 'Empiricism and the Philosophy of Mind', *Minnesota Studies in the Philosophy of Science*, Vol. I, H. Feigl and M. Scriven (eds.), Minneapolis, University of Minnesota Press.
- Schaffner, K.: 1967, 'Approaches to Reduction', *Philosophy of Science*, 34.
- Sharpe, R.: 1987, 'The Very Idea of Folk Psychology,' *Inquiry*, 30.
- Smolensky, P.: 1988, 'On the Proper Treatment of Connectionism', *The Behavioral & Brain Sciences*, 11.
- Stich, S.: 1983, *From Folk Psychology to Cognitive Science*, Cambridge, Mass., Bradford Books/MIT Press.
- Wilkes, K.: 1978, *Physicalism*, London, Routledge and Kegan Paul.

RAMSEY
 Dept. of Philosophy,
 University of Notre Dame

STICH
 Dept. of Philosophy,
 Rutgers University

GARAN
 Dept. of Philosophy,
 University of California, San Diego

ON THE PROPER TREATMENT OF CONNECTIONISM

ABSTRACT: A set of hypotheses is formulated for a connectionist approach to cognitive modeling. These hypotheses are shown to be incompatible with the hypotheses underlying traditional cognitive models. The connectionist models considered are massively parallel numerical computational systems that are a kind of continuous dynamical system. The numerical variables in the system correspond semantically to fine-grained features below the level of the concepts consciously used to describe the task domain. The level of analysis is intermediate between those of symbolic cognitive models and neural models. The explanations of behavior provided are like those traditional in the physical sciences, unlike the explanations provided by symbolic models.

Higher-level analyses of these connectionist models reveal subtle relations to symbolic models. Parallel connectionist memory and linguistic processes are hypothesized to give rise to processes that are describable at a higher level as sequential rule application. At the lower level, computation has the character of massively parallel satisfaction of soft numerical constraints; at the higher level, this can lead to competence characterizable by hard rules. Performance will typically deviate from this competence since behavior is achieved not by interpreting hard rules but by satisfying soft constraints. The result is a picture in which traditional and connectionist theoretical constructs collaborate intimately to provide an understanding of cognition.

KEYWORDS: cognition; connectionism, dynamical systems; networks; neural models; parallel distributed processing; symbolic models

1. INTRODUCTION

In the past half-decade the connectionist approach to cognitive modeling has grown from an obscure cult claiming a few true believers to a movement so vigorous that recent meetings of the Cognitive Science Society have begun to look like connectionist pep rallies. With the rise of the connectionist movement come a number of fundamental questions which are the subject of this target article. I begin with a brief description of connectionist models.

1.1. Connectionist Models

Connectionist models are large networks of simple parallel computing

elements, each of which carries a numerical *activation value* which it computes from the values of neighboring elements in the network, using some simple numerical formula. The network elements, or *units*, influence each other's values through connections that carry a numerical strength, or *weight*. The influence of unit i on unit j is the activation value of unit i times the strength of the connection from i to j . Thus, if a unit has a positive activation value, its influence on a neighbor's value is positive if its weight to that neighbor is positive, and negative if the weight is negative. In an obvious neural allusion, connections carrying positive weights are called *excitatory*, and those carrying negative weights are *inhibitory*.

In a typical connectionist model, input to the system is provided by imposing activation values on the *input units* of the network; these numerical values represent some encoding, or *representation*, of the input. The activation on the input units propagates along the connections until some set of activation values emerges on the *output units*; these activation values encode the output the system has computed from the input. In between the input and output units there may be other units, often called *hidden units*, that participate in representing neither the input nor the output.

The computation performed by the network in transforming the input pattern of activity to the output pattern depends on the set of connection strengths; these weights are usually regarded as encoding the system's knowledge. In this sense, the connection strengths play the role of the program in a conventional computer. Much of the allure of the connectionist approach is that many connectionist networks *program themselves*, that is, they have autonomous procedures for tuning their weights to eventually perform some specific computation. Such learning procedures often depend on training in which the network is presented with sample input/output pairs from the function it is supposed to compute. In learning networks with hidden units, the network itself "decides" what computations the hidden units will perform; because these units represent neither inputs nor outputs, they are never "told" what their values should be, even during training.

In recent years connectionist models have been developed for many tasks, encompassing the areas of vision, language processing, inference, and motor control. Numerous examples can be found in recent proceedings of the meetings of the Cognitive Science Society; *Cognitive Science* (1985); Feldman *et al.* (1985); Hinton and Anderson (1981); McClelland, Rumelhart, and the PDP Research Group (1986); Rumelhart, McClelland, and the PDP Research Group (1986). [See also Ballard 'Cortical Connections and Parallel Processing' *BBS*, 9(1) 1986.]

1.2. Goal of this Target Article

Given the rapid development in recent years of the connectionist approach to cognitive modeling, it is not yet an appropriate time for definitive assessments of the power and validity of the approach. The time seems right, however, for an attempt to articulate the goals of the approach, the fundamental hypotheses it is testing, and the relations presumed to link it with the other theoretical frameworks of cognitive science. A coherent and plausible articulation of these fundamentals is the goal of this target article. Such an articulation is a nontrivial task, because the term "connectionist" encompasses a number of rather disparate theoretical frameworks, all of them quite undeveloped. The connectionist framework I will articulate departs sufficiently radically from traditional approaches in that its relations to other parts of cognitive science are not simple.

For the moment, let me call the formulation of the connectionist approach that I will offer *PTC*. I will not argue the scientific merit of PTC; that some version of connectionism along the lines of PTC constitutes a "proper description of processing" is argued elsewhere (e.g., in Rumelhart, McClelland and the PDP Research Group 1986; McClelland, Rumelhart and the PDP Research Group 1986). Leaving aside the scientific merit of connectionist models, I want to argue here the PTC offers a 'Proper Treatment of Connectionism': a coherent formulation of the connectionist approach that puts it in contact with other theory in cognitive science in a particularly constructive way. PTC is intended as a formulation of connectionism that is at once strong enough to constitute a major cognitive hypothesis, comprehensive enough to face a number of difficult challenges, and sound enough to resist a number of objections in principle. If PTC succeeds in these goals, it will facilitate the real business at hand: Assessing the scientific adequacy of the connectionist approach, that is, determining whether the approach offers computational power adequate for human cognitive competence and appropriate computational mechanisms to accurately model human cognitive performance.

PTC is a response to a number of positions that are being adopted concerning connectionism – pro, con, and blandly ecumenical. These positions, which are frequently expressed orally but rarely set down in print, represent, I believe failures of supporters and critics of the traditional approach truly to come to grips with each other's views. Advocates of the traditional approach to cognitive modeling and AI (artificial intelligence) are often willing to grant that connectionist systems are useful, perhaps even important, for modeling lower-level processes (e.g. early vision), or for fast an fault-tolerant im-

plementation of conventional AI programs, or for understanding how the brain might happen to implement LISP. These ecumenical positions, I believe, fail to acknowledge the true challenge that connectionists are posing to the received view of cognition; PTC is an explicit formulation of this challenge.

Other supporters of the traditional approach find the connectionist approach to be fatally flawed because it cannot offer anything new (since Universal Turing machines are, after all, “universal”), or because it cannot offer the kinds of explanations that cognitive science requires. Some dismiss connectionist models on the grounds that they are too neurally unfaithful. PTC has been designed to withstand these attacks.

On the opposite side, most existing connectionist models fail to come to grips with the traditional approach – partly through a neglect intended as benign. It is easy to read into the connectionist literature the claim that there is no role in cognitive science for traditional theoretical constructs such as rules, sequential processing, logic, rationality, and conceptual schemata or frames. PTC undertakes to assign these constructs their proper role in a connectionist paradigm for cognitive modeling. PTC also addresses certain foundational issues concerning mental states.

I see no way of achieving the goals of PTC without adopting certain positions that will be regarded by a number of connectionists as premature or mistaken. These are inevitable consequences of the fact that the connectionist approach is still quite underdeveloped, and that the term “connectionist” has come to label a number of approaches that embody significantly conflicting assumptions. PTC is *not* intended to represent a consensus view of what the connectionist approach is or should be.

It will perhaps enhance the clarity of the article if I attempt at the outset to make my position clear on the present value of connectionist models and their future potential. This article is not intended as a defense of all these views, though I will argue for a number of them, and the remainder have undoubtedly influenced the presentation. On the one hand, I believe that:

- (1) a. It is far from clear whether connectionist models have adequate computational power to perform high-level cognitive tasks: There are serious obstacles that must be overcome before connectionist computation can offer modelers power comparable to that of symbolic computation.
- b. It is far from clear that connectionist models offer a sound basis for modeling human cognitive performance: The connectionist

approach is quite difficult to put into detailed contact with empirical methodologies.

- c. It is far from clear that connectionist models can contribute to the study of human competence: Connectionist models are quite difficult to analyze for the kind of highlevel properties required to inform the study of human competence.
- d. It is far from clear that connectionist models, in something like their present forms, can offer a sound basis for modeling neural computation: As will be explicitly addressed in Section 4, there are many serious gaps between the connectionist models and current views of important neural properties.
- e. Even under the most successful scenario for connectionist cognitive science, many of the currently practiced research strategies in cognitive science would remain viable and productive.

On the other hand, I believe that:

- (1)
 - f. It is very likely that the connectionist approach will contribute significant, long-lasting ideas to the rather impoverished theoretical repertoire of cognitive science.
 - g. It is very likely that connectionist models will turn out to offer contributions to the modeling of human cognitive performance on higher-level tasks that are at least as significant as those offered by traditional, symbolic, models.
 - h. It is likely that the view of the competence/performance distinction that arises from the connectionist approach will successfully heal a deep and ancient rift in the science and philosophy of mind.
 - i. It is likely that connectionist models will offer the most significant progress of the past several millennia on the mind/body problem.
 - j. It is very likely that, given the impoverished theoretical repertoire of computational neuroscience, connectionist models will serve as an excellent stimulus to the development of models of neural computation that are significantly better than both current connectionist models and current neural models.
 - k. There is a reasonable chance that connectionist models will lead to the development of new somewhat-general-purpose self-programming, massively parallel analog computers, and a new

theory of analog parallel computation: They may possibly even challenge the strong construal of Church's Thesis as the claim that the class of well-defined computations is exhausted by those of Turing machines.

1.3. Levels of Analysis

Most of the foundational issues surrounding the connectionist approach turn, in one way or another, on the level of analysis adopted. The terminology, graphics, and discussion found in most connectionist papers strongly suggest that connectionist modeling operates at the neural level. I will argue, however, that it is better *not* to construe the principles of cognition being explored in the connectionist approach as the principles of the neural level. Specification of the level of cognitive analysis adopted by PTC is a subtle matter which consumes much of this article. To be sure, the level of analysis adopted by PTC is lower than that of the traditional, symbolic paradigm; but, at least for the present, the level of PTC is more explicitly related to the level of the symbolic paradigm than it is to the neural level. For this reason I will call the paradigm for cognitive modeling proposed by PTC the *subsymbolic paradigm*.

A few comments on terminology. I will refer to the traditional approach to cognitive modeling as the *symbolic paradigm*. Note that I will always use the term "symbolic paradigm" to refer to the traditional approach to cognitive *modeling*: the development of AI-like computer programs to serve as models of psychological performance. The symbolic paradigm in cognitive modeling has been articulated and defended by Newell and Simon (1976; Newell 1980), as well as by Fodor (1975; 1987), Pylyshyn (1984), and others. The fundamental hypotheses of this paradigm embrace most of mainstream AI, in addition to AI-based systems that are explicitly offered as models of human performance. The term "symbolic paradigm" is explicitly *not* intended to encompass competence theories such as the formal theory of grammar; such competence theories bear deep relations to the symbolic paradigm but they are not a focus of attention in this paper. In particular, much of the work in formal linguistics differs from the symbolic paradigm in cognitive modeling in many of the same ways as the connectionist approach I will consider; on a number of the dimensions I will use to divide the symbolic and subsymbolic paradigms, much linguistics research falls on the subsymbolic side.

I have found it necessary to deal only with a subset of the symbolic and connectionist approaches in order to get beyond superficial, syntactic

issues. On the symbolic side, I am limiting consideration to the Newell/Simon/Fodor/Polyshyn view of cognition, and excluding, for example, the view adopted by much of linguistics; on the connectionist side, I will consider only a particular view, the “subsymbolic paradigm,” and exclude a number of competing connectionist perspectives. The only alternative I see at this point is to characterize the symbolic and connectionist perspectives so diffusely that substantive analysis becomes impossible.

In calling the traditional approach to cognitive modeling the “symbolic paradigm,” I intend to emphasize that in this approach, cognitive descriptions are built of entities that are symbols both in the semantic sense of referring to external objects and in the syntactic sense of being operated upon by symbol manipulation. These manipulations model fundamental psychological processes in this approach to cognitive modeling.

The name “subsymbolic paradigm” is intended to suggest cognitive descriptions built up of entities that correspond to *constituents* of the symbols used in the symbolic paradigm; these fine-grained constituents could be called *subsymbols*, and they are the activities of individual processing units in connectionist networks. Entities that are typically represented in the symbolic paradigm by symbols are typically represented in the subsymbolic paradigm by a large number of subsymbols. Along with this semantic distinction comes a syntactic distinction. Subsymbols are not operated upon by symbol manipulation: They participate in numerical – not symbolic – computation. Operations in the symbolic paradigm that consist of a single discrete operation (e.g., a memory fetch) are often achieved in the subsymbolic paradigm as the result of a large number of much finer-grained (numerical) operations.

Since the level of cognitive analysis adopted by the subsymbolic paradigm for formulating connectionist models is lower than the level traditionally adopted by the symbolic paradigm, for the purposes of relating these two paradigms, it is often important to analyze connectionist models at a higher level; to amalgamate, so to speak, the subsymbols into symbols. Although the symbolic and subsymbolic paradigms each have their preferred level of analysis, the cognitive models they offer can be described at multiple levels. It is therefore useful to have distinct names for the levels: I will call the preferred level of the symbolic paradigm the *conceptual level* and that of the subsymbolic paradigm the *subconceptual level*. These names are not ideal, but will be further motivated in the course of characterizing the levels. A primary goal of this article is to articulate a coherent set of hypotheses about the subconceptual level: the kind of cognitive descriptions that are used, the

computational principles that apply, and the relations between the subconceptual and both the symbolic and neural levels.

The choice of level greatly constrains the appropriate formalism for analysis. Probably the most striking feature of the connectionist approach is the change in formalism relative to the symbolic paradigm. Since the birth of cognitive science, *language* has provided the dominant theoretical model. Formal cognitive models have taken their structure from the syntax of formal languages, and their content from the semantics of natural language. The mind has been taken to be a machine for formal symbol manipulation, and the symbols manipulated have assumed essentially the same semantics as words of English.

The subsymbolic paradigm challenges both the syntactic and semantic role of language in formal cognitive models. Section 2 formulates this challenge. Alternative fillers are described for the roles language has traditionally played in cognitive science, and the new role left to language is delimited. The fundamental hypotheses defining the subsymbolic paradigm are formulated, and the challenge that nothing new is being offered is considered. Section 4 considers the relation between the subsymbolic paradigm and neuroscience; the challenge that connectionist models are too neurally unfaithful is addressed. Section 5 presents the relations between analyses of cognition at the neural, subconceptual, and conceptual levels. It also previews the remainder of the article, which deals with the relations between the subconceptual and conceptual levels; the types of explanations of behavior provided by the symbolic and subsymbolic paradigms are then discussed. Section 6 faces the challenge of accounting for conscious, rule-guided behavior within the subsymbolic paradigm. Section 7 addresses the challenge of distinguishing cognitive from noncognitive systems at the subconceptual level. Various properties of subsymbolic mental states, and the issue of rationality, are considered. Section 8 elaborates briefly on the computational principles that apply the subconceptual level. Section 9 discusses how higher, conceptual-level descriptions of subsymbolic models approximate symbolic models (under their conceptual-level descriptions).

In this target article I have tried to typographically isolate concise formulations of the main points. Most of these numbered points serve to characterize the subsymbolic paradigm, but a few define alternative points of view; to avoid confusion, the latter have been explicitly tagged by the phrase, *To be rejected*.

2. FORMALIZATION OF KNOWLEDGE

2.1 *Cultural Knowledge and Conscious Rule Interpretation*

What is an appropriate formalization of the knowledge that cognitive agents possess and the means by which they use that knowledge to perform cognitive tasks? As a starting point, we can look to those knowledge formalizations that predate cognitive science. The most formalized knowledge is found in sciences like physics that rest on mathematical principles. Domain knowledge is formalized in linguistic structures such as “energy is conserved” (or an appropriate encryption), and logic formalizes the use of that knowledge to draw conclusions. Knowledge consists of axioms, and drawing conclusions consists of proving theorems.

This method of formulating knowledge and drawing conclusions has extremely valuable properties.

- (2) a. *Public access*: The knowledge is accessible to many people.
- b. *Reliability*: Different people (or the same person at different times) can reliably check whether conclusions have been validly reached.
- c. *Formality, bootstrapping, universality*: The inferential operations require very little experience with the domain to which the symbols refer.

These three properties are important for science because it is a cultural activity. It is of limited social value to have knowledge that resides purely in one individual (2a). It is of questionable social value to have knowledge formulated in such a way that different users draw different conclusions (e.g., can't agree that an experiment falsifies a theory) (2b). For cultural propagation of knowledge, it is helpful if novices with little or no experience with a task can be given a means for performing that task, and thereby a means for acquiring experience (2c).

There are cultural activities other than science that have similar requirements. The laws of a nation and the rules of an organization are also linguistically formalized procedures for effecting action which different people can carry out with reasonable reliability. In all these cases, the goal is to create an abstract decision system that resides outside any single person.

Thus, at the cultural level, the goal is to express knowledge in a form that can be executed reliably by different people, even inexperienced ones. We can view the top-level conscious processor of individual people as a *virtual*

machine – the *conscious rule interpreter* – and we can view cultural knowledge as a program that runs on that machine. Linguistic formulations of knowledge are perfect for this purpose. The procedures that different people can reliably execute are explicit, step-by-step linguistic instructions. This is what has been formalized in the theory of *effective procedures* (Turing 1936). Thanks to property (2c), the top-level conscious human processor can be idealized as universal: capable of executing any effective procedure. The theory of effective procedures – the classical theory of computation (Hopcroft and Ullman, 1979) – is physically manifest in the von Neumann (serial) computer. One can say that the von Neumann computer is a machine for automatically following the kinds of explicit instructions that people can fairly reliably follow – but much faster and with perfect reliability.

Thus we can understand why the production system of computation theory, or more generally the von Neumann computer, has provided a successful model of how people execute instructions (e.g., models of novice physics problem solving such as that of Larkin *et al.* 1980). In short, when people (e.g., novices) consciously and sequentially follow rules (such as those they have been taught), their cognitive processing is naturally modeled as the sequential interpretation¹ of a linguistically formalized procedure. The rules being followed are expressed in terms of the consciously accessible concepts with which the task domain is conceptualized. In this sense, the rules are formulated at the conceptual level of analysis.

To sum up:

- (3)
 - a. Rules formulated in natural language can provide an effective formalization of cultural knowledge.
 - b. Conscious rule application can be modeled as the sequential interpretation of such rules by a virtual machine called the conscious rule interpreter.
 - c. These rules are formulated in terms of the concepts consciously used to describe the task domain – they are formulated at the conceptual level.

2.2. *Individual Knowledge, Skill, and Intuition in the Symbolic Paradigm*

The constraints on cultural knowledge formalization are not the same as those on individual knowledge formalization. The intuitive knowledge in a physics expert or a native speaker may demand, for a truly accurate description, a

formalism that is not a good one for cultural purposes. After all, the individual knowledge in an expert's head does not possess the properties (2) of cultural knowledge: It is not publically accessible or completely reliable, and it is completely dependent on ample experience. Individual knowledge is a program that runs on a virtual machine that need not be the same as the top-level conscious processor that runs the cultural knowledge. By definition, conclusions reached by intuition do not come from conscious application of rules, and intuitive processing need not have the same character as conscious rule application.

What kinds of programs are responsible for behavior that is not conscious rule application? I will refer to the virtual machine that runs these programs as the *intuitive processor*. It is presumably responsible for all of animal behavior and a huge portion of human behavior: Perception, practiced motor behavior, fluent linguistic behavior, intuition in problem solving and game playing – in short, practically all skilled performance. The transference of responsibility from the conscious rule interpreter to the intuitive processor during the acquisition of skill is one of the most striking and well-studied phenomena in cognitive science (Anderson 1981). An analysis of the formalization of knowledge must consider both the knowledge involved in novices' conscious application of rules and the knowledge resident in experts' intuition, as well as their relationship.

An appealing possibility is this:

- (4) a. The programs running on the intuitive processor consist of linguistically formalized rules that are sequentially interpreted.
(*To be rejected.*)

This has traditionally been the assumption of cognitive science. Native speakers are unconsciously interpreting rules, as are physics experts when they are intuiting answers to problems. Artificial intelligence systems for natural language processing and problem solving are programs written in a formal language for the symbolic description of procedures for manipulating symbols.

To the syntactic hypothesis (4a) a semantic one corresponds:

- (4) b. The programs running on the intuitive processor are composed of elements, that is, symbols, referring to essentially the same concepts as the ones used to consciously conceptualize the task domain. (*To be rejected.*)

This applies to production system models in which the productions represent-

ing expert knowledge are compiled versions of those of the novice (Anderson 1983; Lewis 1978) and to the bulk of AI programs.

Hypotheses (4a) and (4b) together comprise:

- (4) The unconscious rule interpretation hypothesis: (*To be rejected.*)
The programs running on the intuitive processor have a syntax and semantics comparable to those running on the conscious rule interpreter.

This hypothesis has provided the foundation for the symbolic paradigm for cognitive modeling. Cognitive models of both conscious rule application and intuitive processing have been programs constructed of entities which are *symbols* both in the syntactic sense of being operated on by symbol manipulation and in the semantic sense of (4b). Because these symbols have the conceptual semantics of (4b), I am calling the level of analysis at which these programs provide cognitive models the *conceptual level*.

2.3. *The Subsymbolic Paradigm and Intuition*

The hypothesis of unconscious rule interpretation (4) is an attractive possibility which a connectionist approach to cognitive modeling rejects. Since my purpose here is to formulate rather than argue the scientific merits of a connectionist approach, I will not argue against (4) here. I will point out only that in general, connectionists do not casually reject (4). Several of today's leading connectionist researchers were intimately involved with serious and longstanding attempts to make (4) serve the needs of cognitive science.² Connectionists tend to reject (4) because they find the consequences that have actually resulted from its acceptance to be quite unsatisfactory, for a number of quite independent reasons, including:

- (5) a. Actual AI systems built on hypothesis (4) seem too brittle, to inflexible, to model true human expertise.
b. The process of articulating expert knowledge in rules seems impractical for many important domains (e.g., common sense).
c. Hypothesis (4) has contributed essentially no insight into how knowledge is represented in the brain.

What motivates the pursuit of connectionist alternatives to (4) is a hunch that such alternatives will better serve the goals of cognitive science. Substantial empirical assessment of this hunch is probably at least a decade away. One possible alternative to (4a) is:

- (6) The neural architecture hypothesis: (*To be rejected.*) The intuitive processor for a particular task uses the same architecture that the brain uses for that task.

Whatever appeal this hypothesis might have, it seems incapable in practice of supporting the needs of the vast majority of cognitive models. We simply do not know what architecture the brain uses for performing most cognitive tasks. There may be some exceptions (such as visual and spatial tasks), but for problem solving, language, and many others (6) simply cannot do the necessary work at the present time.

These points and others relating to the neural level will be considered in more detail in Section 4. For now the point is simply that characterizing the level of analysis of connectionist modeling is not a matter of simply identifying it with the neural level. While the level of analysis adopted by most connectionist cognitive models is not the conceptual one, it is also not the neural level. [See also Anderson: 'Methodologies for Studying Human Knowledge', *BBS*, 10(3), 1987.]

The goal now is to formulate a connectionist alternative to (4) that, unlike (6), provides a viable basis for cognitive modeling. A first, crude approximation to this hypothesis is:

- (7) The intuitive processor has a certain kind of connectionist architecture (which abstractly models a few of the most general features of neural networks). (*To be elaborated.*)

Postponing consideration of the neural issues to Section 4, we now consider the relevant kind of connectionist architecture.

The view of the connectionist architecture I will adopt is the following (for further treatment of this viewpoint, see Smolensky 1986b). The numerical activity of all the processors in the network form a large *state vector*. The interactions of the processors, the equations governing how the activity vector changes over time as processors respond to one another's values, is an *activation evolution equation*. This evolution equation governing the mutual interactions of the processors involves the connection weights: numerical parameters which determine the direction and magnitude of the influence of one activation value on another. The activation equation is a differential equation (usually approximated by the finite difference equation that arises from discrete time slices; the issue of discrete approximation is taken up in Section 8.1). In learning systems, the connection weights change during training according to the learning rule, which is another differential equation:

the *connection evolution equation*.

Knowledge in a connectionist system lies in its connection strengths. Thus, for the first part of our elaboration on (7) we have the following alternative to (4a):

- (8) a. The connectionist dynamical hypothesis: The state of the intuitive processor at any moment is precisely defined by a vector of numerical values (one for each unit). The dynamics of the intuitive processor are governed by a differential equation. The numerical parameters in this equation constitute the processor's program or knowledge. In learning systems, these parameters change according to another differential equation.

This hypothesis states that the intuitive processor is a certain kind of *dynamical system*: Like the dynamical systems traditionally studied in physics, the state of the system is a numerical vector evolving in time according to differential evolution equations. The special properties that distinguish this kind of dynamical system – a *connectionist dynamical system* – are only vaguely described in (8a). A much more precise specification is needed. It is premature at this point to commit oneself to such a specification, but one large class of subsymbolic models is that of quasilinear dynamical systems, explicitly discussed in Smolensky (1986b) and Rumelhart, Hinton, and Williams (1986). Each unit in a quasilinear system computes its value by first calculating the weighted sum of its inputs from other units and then transforming this sum with a nonlinear function. An important goal of the subsymbolic paradigm is to characterize the computational properties of various kinds of connectionist dynamical systems (such as quasilinear systems) and thereby determine which kinds provide appropriate models of various types of cognitive processes.

The connectionist dynamical system hypothesis (8a) provides a connectionist alternative to the syntactic hypothesis (4a) of the symbolic paradigm. We now need a semantic hypothesis compatible with (8a) to replace (4b). The question is: What does a unit's value *mean*? The most straightforward possibility is that the semantics of each unit is comparable to that of a word in natural language; each unit represents such a concept, and the connection strengths between units reflect the degree of association between the concepts.

- (9) The conceptual unit hypothesis: (*To be rejected.*) Individual intuitive processor elements – individual units – have essentially

the same semantics as the conscious rule interpreter's elements, namely, words of natural language.

But (8a) and (9) make an infertile couple. Activation of concepts spreading along degree of association links may be adequate for modeling simple aspects of cognition – such as relative times for naming words or the relative probabilities of perceiving letters in various contexts – but it cannot be adequate for complex tasks such as question answering or grammatical judgments. The relevant structures cannot even be feasibly represented in such a network, let alone effectively processed.

Great computational power must be present in the intuitive processor to deal with the many cognitive processes that are extremely complex when described at the conceptual level. The symbolic paradigm, based on hypothesis (4), gets its power by allowing highly complex, essentially arbitrary, operations on symbols with conceptual-level semantics: simple semantics, complex operations. If the operations are required to be as simple as those allowed by hypothesis (8a), we cannot get away with a semantics as simple as that of (9).³ A semantics compatible with (8a) must be more complicated:

- (8) b. The subconceptual unit hypothesis: The entities in the intuitive processor with the semantics of conscious concepts of the task domain are complex patterns of activity over many units. Each unit participates in many such patterns.

(See several of the papers in Hinton and Anderson 1981; Hinton, McClelland and Rumelhart 1986; the neural counterpart is associated with Hebb 1949; Lashley 1950, about which see Feldman 1986.) The interactions between *individual units* are simple, but these units do not have conceptual semantics: they are *subconceptual*. The interactions between the entities with conceptual semantics, interactions between complex patterns of activity, are not at all simple. Interactions at the level of activity patterns are not directly described by the formal definition of a subsymbolic model; they must be computed by the analyst. Typically, these interactions can be computed only approximately. In other words, there will generally be no precisely valid, complete, computable formal principles at the conceptual level; such principles exist only at the level of individual units – the *subconceptual level*.

- (8) c. The subconceptual level hypothesis: Complete, formal, and precise descriptions of the intuitive processor are generally

tractable not at the conceptual level, but only at the subconceptual level.

In (8c), the qualification “complete, formal, and precise” is important: Conceptual-level descriptions of the intuitive processor’s performance can be derived from the subconceptual description, but, unlike the description at the subconceptual level, the conceptual-level descriptions will be either incomplete (describing only certain aspects of the processing) or informal (describing complex behaviors in, say, qualitative terms) or imprecise (describing the performance up to certain approximations or idealizations such as “competence” idealizations away from actual performance). Explicit examples of each of these kinds of conceptual-level descriptions of subsymbolic systems will be considered in Section 9.

Hypotheses (8a–c) can be summarized as:

- (8) The subsymbolic hypothesis:
The intuitive processor is a subconceptual connectionist dynamical system that does not admit a complete, formal, and precise conceptual-level description.

This hypothesis is the cornerstone of the subsymbolic paradigm.⁴

2.4. The Incompatibility of the Symbolic and Subsymbolic Paradigms

I will now show that the symbolic and subsymbolic paradigms, as formulated above, are incompatible – that hypotheses (4) and (8) about the syntax and semantics of the intuitive processor are not mutually consistent. This issue requires care, because it is well known that one virtual machine can often be implemented in another, that a program written for one machine can be translated into a program for the other. The attempt to distinguish subsymbolic and symbolic computation might well be futile if each can simulate the other. After all, a digital computer is in reality some sort of dynamical system simulating a von Neumann automaton, and in turn, digital computers are usually used to simulate connectionist models. Thus it seems possible that the symbolic and subsymbolic hypotheses (4) and (8) are *both* correct: The intuitive processor can be regarded as a virtual machine for sequentially interpreting rules on one level *and* as a connectionist machine on a lower level.

This possibility fits comfortably within the symbolic paradigm, under a formulation such as:

- (10) Valid connectionist models are merely implementations, for a certain kind of parallel hardware, of symbolic programs that provide exact and complete accounts of behavior at the conceptual level. (*To be rejected.*)

However (10) contradicts hypothesis (8c), and is thus incompatible with the subsymbolic paradigm. The symbolic programs that (4) hypothesizes for the intuitive processor could indeed be translated for a connectionist machine; but the translated programs would *not* be the kind of subsymbolic program that (8) hypothesizes. If (10) is correct, (8) is wrong; at the very least, (8c) would have to be removed from the defining hypothesis of the subsymbolic paradigm, weakening it to the point that connectionist modeling does become mere implementation. Such an outcome would constitute a genuine defeat of a research program that I believe many connectionists are pursuing.

What about the reverse relationship, where a symbolic program is used to implement a subsymbolic system? Here it is crucial to realize that the symbols in such programs represent the activation values of units and the strengths of connections. By hypothesis (8b), these do not have conceptual semantics, and thus hypothesis (4b) is violated. The subsymbolic programs that (8) hypothesizes for the intuitive processor can be translated for a von Neumann machine, but the translated programs are *not* the kind of symbolic program that (4) hypothesizes.

These arguments show that unless the hypotheses of the symbolic and subsymbolic paradigms are formulated with some care, the substance of the scientific issue at stake can easily be missed. It is well known that von Neumann machines and connectionist networks can simulate each other. This fact leads some people to adopt the position that the connectionist approach cannot offer anything fundamentally new because we already have Turing machines and, following Church's Thesis, reason to believe that, when it comes to computation, Turing machines are everything. This position, however, mistakes the issue for cognitive science to be the purely syntactic question of whether mental programs are written for Turing/von Neumann machines or connectionist machines. This is a nonissue. If one cavalierly characterizes the two approaches *only syntactically*, using (4a) and (8a) alone, then indeed the issue – connectionist or not connectionist – appears to be “one of AI's wonderful red herrings.”⁵

It is a mistake to claim that the connectionist approach has nothing new to offer cognitive science. The issue at stake is a central one: Does the complete formal account of cognition lie at the conceptual level? The position taken by the subsymbolic paradigm is: No – it lies at the subconceptual level.

3. REPRESENTATION AT THE SUBCONCEPTUAL LEVEL

Having hypothesized the existence of a subconceptual level, we must now consider its nature. Hypothesis (8b) leaves open important questions about the semantics of subsymbolic systems. What kind of subconceptual features do the units in the intuitive processor represent? Which activity patterns actually correspond to particular concepts or elements of the problem domain?

There are no systematic or general answers to these questions at the present time; seeking answers is one of the principal tasks for the subsymbolic research paradigm. At present, each individual subsymbolic model adopts particular procedures for relating patterns of activity – activity vectors – to the conceptual-level descriptions of inputs and outputs that define the model's task. The vectors chosen are often values of fine-grained features of the inputs and outputs, based on some preexisting theoretical analysis of the domain. For example, for the task studied by Rumelhart and McClelland (1986), transforming root phonetic forms of English verbs to their past-tense forms, the input and output phonetic strings are represented as vectors of values for context-dependent binary phonetic features. The task description at the conceptual level involves consciously available concepts such as the words "go" and "went," while the subconceptual level used by the model involves a very large number of fine-grained features such as "roundedness preceded by frontality and followed by backness." The representation of "go" is a large pattern of activity over these features.

Substantive progress in subsymbolic cognitive science requires that systematic commitments be made to vectorial representations for individual cognitive domains. It is important to develop mathematical or empirical methodologies that can adequately constrain these commitments. The vectors chosen to represent inputs and outputs crucially affect a model's predictions, since the generalizations the model makes are largely determined by the similarity structure of the chosen vectors. Unlike symbolic tokens, these vectors lie in a topological space in which some are close together and others far apart.

What kinds of methodologies might be used to constrain the representation at the subconceptual level? The methodology used by Rumelhart and McClelland (1986) in the past-tense model is one that has been fairly widely practiced, particularly in models of language processing: Representational features are borrowed from existing theoretical analyses of the domain and adapted (generally in somewhat ad hoc ways) to meet the needs of connec-

tionist modeling. This methodology clearly renders the subsymbolic approach dependent on other research paradigms in the cognitive sciences and suggests that, certainly in the short term, the subsymbolic paradigm cannot *replace* these other research paradigms. (This is a theme I will return to in the conclusion of the paper.)

A second possible theoretical methodology for studying subconceptual representation relates to the learning procedures that can train hidden units in connectionist networks. Hidden units support internal representations of elements of the problem domain, and networks that train their hidden units are in effect learning effective subconceptual representations of the domain. If we can analyze the representations that such networks develop, we can perhaps obtain principles of subconceptual representation for various problem domains.

A third class of methodology views the task of constraining subconceptual models as the calibration of connectionist models to the human cognitive system. The problem is to determine what vectors should be assigned to represent various aspects of the domain so that the resulting behavior of the connectionist model matches human behavior. Powerful mathematical tools are needed for relating the overall behavior of the network to the choice of representational vectors; ideally, these tools should allow us to *invert* the mapping from representations to behavior so that by starting with a mass of data on human performance we can turn a mathematical crank and have representational vectors pop out. An example of this general type of tool is the technique of *multidimensional scaling* (Shepard 1962), which allows data on human judgments of the similarity between pairs of items in some set to be turned into vectors for representing those items (in a sense). The subsymbolic paradigm needs tools such as a version of multidimensional scaling based on a connectionist model of the process of producing similarity judgments.

Each of these methodologies poses serious research challenges. Most of these challenges are currently being pursued, so far with at best modest success. In the first approach, systematic principles must be developed for adapting to the connectionist context the featural analyses of domains that have emerged from traditional, nonconnectionist paradigms. These principles must reflect fundamental properties of connectionist computation, for otherwise, the hypothesis of connectionist computation is doing no work in the study of mental representation. In the second methodology, principles must be discovered for the representations learned by hidden units, and in the third methodology, principles must be worked out for relating choices of

representational vectors to overall system behavior. These are challenging mathematical problems on which the ultimate success of the subsymbolic paradigm rests. Sections 8 and 9 discuss some results related to these mathematical problems, but they are far from strong enough to carry the necessary weight.

The next two sections discuss the relation between the subconceptual level and other levels: The relation to the neural levels is addressed in Section 4, and the relation to the conceptual level is taken up in Section 5.

4. THE SUBCONCEPTUAL AND NEURAL LEVELS

The discussion in the preceding section overlooks an obvious methodology for constraining subconceptual representations – just look at how the brain does it. This brings us back to the parenthetical comment in (7) and the general issue of the relation between the subconceptual and neural levels.⁶

The relation between the subconceptual and neural levels can be addressed in both syntactic and semantic terms. The semantic question is the one just raised: How do representations of cognitive domains as patterns of activity over subconceptual units in the network models of the subsymbolic paradigm relate to representations over neurons in the brain? The syntactic question is: How does the processing architecture adopted by networks in the subsymbolic paradigm relate to the processing architecture of the brain?

There is not really much to say about the semantic question because so little is known about neural representation of higher cognitive domains. When it comes to connectionist modeling of say, language processing, the “just look at how the brain does it” methodology doesn’t take one very far towards the goal of constructing a network that does the task at all. Thus it is unavoidable that, for the time being, in subsymbolic models of higher processes, the semantics of network units are much more directly related to conceptual level accounts of these processes than to any neural account. Semantically, the subconceptual level seems at present rather close to the conceptual level, while we have little ground for believing it to be close to the neural level.

This conclusion is at odds with the commonly held view that connectionist models are neural models. That view presumably reflects a bias against semantic considerations in favor of syntactic ones. If one looks only at processing mechanisms, the computation performed by subsymbolic models seems much closer to that of the brain than to that of symbolic models. This suggests that syntactically, the subconceptual level is closer to the neural level than to the conceptual level.

TABLE I
Relations between the neural and subsymbolic architectures

<i>Cerebral cortex</i>		<i>Connectionist dynamical systems</i>
State defined by continuous numerical variables (potentials, synaptic areas,...)	+	State defined by continuous numerical variables (activations, connection strengths)
State variables change continuously in time	+	State variables change continuously in time
Interneuron interaction parameters changeable; seat of knowledge	+	Interunit interaction parameters changeable; seat of knowledge
Huge number of state variables	+	Large number of state variables
High interactional complexity (highly nonhomogeneous interactions)	+	High interactional complexity (highly nonhomogeneous interactions)
Neurons located in 2+1-d space have dense connectivity to nearby neurons; have geometrically mapped connectivity to distant neurons	–	Units have no spatial location uniformly dense connections
Synapses located in 3-d space; locations strongly affect signal interactions	–	Connections have no spatial location
Distal projections between areas have intricate topology	–	Distal projections between node pools have simple topology
Distal interactions mediated by discrete signals	–	All interactions non-discrete
Intricate signal integration at single neuron	–	Signal integration is linear
Numerous signal types	–	Single signal type

Let us take then the syntactic question: Is the processing architecture adopted by subsymbolic models (8a) well-suited for describing processing at the neural level? Table I presents some of the relations between the architectures. The left column lists currently plausible features of some of the most general aspects of the neural architecture, considered at the level of neurons (Crick and Asanuma 1986). The right column lists the corresponding architectural features of the connectionist dynamical systems typically used in subsymbolic models. In the center column, each hit has been indicated by a + and each miss by a -.

In Table I the loose correspondence assumed is between neurons and units, between synapses and connections. It is not clear how to make this correspondence precise. Does the activity of a unit correspond to the membrane potential at the cell body? Or the time-averaged firing rate of the neuron? Or the population-averaged firing rate of many neurons? Since the integration of signals between dendritic trees is probably more like the linear integration appearing in quasilinear dynamical systems than is the integration of synaptic signals on a dendrite, would it not be better to view a connection not as an individual synaptic contact but rather as an aggregate contact on an entire dendritic tree?

Given the difficulty of precisely stating the neural counterpart of components of subsymbolic models, and given the significant number of misses, even in the very general properties considered in Table I, it seems advisable to keep the question open of the detailed relation between cognitive descriptions at the subconceptual and neural levels. There seems no denying, however, that the subconceptual level is significantly closer to the neural level than is the conceptual level: Symbolic models possess even fewer similarities with the brain than those indicated in Table I.

The subconceptual level ignores a great number of features of the neural level that are probably extremely important to understanding how the brain computes. Nonetheless, the subconceptual level does incorporate a number of features of neural computation that are almost certainly extremely important to understanding how the brain computes. The general principles of computation at the subconceptual level – computation in high-dimensional, high-complexity dynamical systems – *must* apply to computation in the brain; these principles are likely to be necessary, if not sufficient, to understand neural computation. And while subconceptual principles are not unambiguously and immediately applicable to neural systems, they are certainly more readily applicable than the principles of symbolic computation.

In sum:

- (11) The fundamental level of the subsymbolic paradigm, the subconceptual level, lies between the neural and conceptual levels.

As stated earlier, on semantic measures, the subsymbolic level seems closer to the conceptual level, whereas on syntactic measures, it seems closer to the neural level. It remains to be seen whether, as the subsymbolic paradigm develops, this situation will sort itself out. Mathematical techniques like those discussed in the previous section may yield insights into subsymbolic representation that will increase the semantic distance between the subconceptual and conceptual levels. There are already significant indications that as new insights into subsymbolic computation are emerging, and additional information processing power is being added to subsymbolic models, the syntactic distance between the subconceptual and neural levels is increasing. In the drive for more computational power, architectural decisions seem to be driven more and more by mathematical considerations and less and less by neural ones.⁷

Once (11) is accepted, the proper place of subsymbolic models in cognitive science will be clarified. It is common to hear dismissals of a particular subsymbolic model because it is not immediately apparent how to implement it precisely in neural hardware, or because certain neural features are absent from the model. We can now identify two fallacies in such a dismissal. First, following (11): Subsymbolic models should not be viewed as neural models. If the subsymbolic paradigm proves valid, the best subsymbolic models of a cognitive process should one day be shown to be some reasonable higher-level approximation to the neural system supporting that process. This provides a heuristic that favors subsymbolic models that seem more likely to be reducible to the neural level. But this heuristic is an extremely weak one given how difficult such a judgment must be with the current confusion about the precise neural correlates of units and connections, and the current state of both empirical and theoretical neuroscience.

The second fallacy in dismissing a particular subsymbolic model because of neural unfaithfulness rests on a failure to recognize the role of individual models in the subsymbolic paradigm. A model can make a valuable contribution by providing evidence for general principles that are characteristic of a broad class of subsymbolic systems. The potential value of "ablation" studies of the NETtalk text-to-speech system (Sejnowski and Rosenberg 1986), for example, does not depend entirely on the neural faithfulness of the model, or

even on its psychological faithfulness. NETtalk is a subsymbolic system that performs a complex task. What happens to its performance when internal parts are damaged? This provides a significant clue to the general principles of degradation in *all* complex subsymbolic systems: Principles that will apply to future systems that are more faithful as models.

There are, of course, many neural models that do take many of the constraints of neural organization seriously, and for which the analogue of Table I would show nearly all hits. But we are concerned here with connectionist models for performing cognitive tasks, and these models typically possess the features displayed in Table I, with perhaps one or two deviations. The claim is not that neural models don't exist, but rather that they should not be confused with subsymbolic models.

Why is it that neural models of cognitive processes are, generally speaking, currently not feasible? The problem is not an insufficient quantity of data about the brain. The problem, it seems, is that the data are generally of the wrong kind for cognitive modeling. Our information about the nervous system tends to describe its structure, not its dynamic behavior. Subsymbolic systems are dynamical systems with certain kinds of differential equations governing their dynamics. If we knew which dynamical variables in the neural system for some cognitive task were the critical ones for performing that task, and what the "equations of motion" were for those variables, we could use that information to build neurally faithful cognitive models. But generally what we know instead are endless static properties of how the hardware is arranged. Without knowing which (if any) of these structures support relevant dynamical processes, and what equations govern those processes, we are in a position comparable to someone attempting to model the solar system, armed with voluminous data on the colored bands of the planets but with no knowledge of Newton's Laws.

To summarize:

- (12) a. Unlike the symbolic architecture, the subsymbolic architecture possesses a number of the most general features of the neural architecture.
- b. However, the subsymbolic architecture lacks a number of the more detailed but still quite general features of the neural architecture; the subconceptual level of analysis is higher than the neural level.
- c. For most cognitive functions, neuroscience cannot provide the

relevant information to specify a cognitive model at the neural level.

- d. The general cognitive principles of the subconceptual level will probably be important contributors to future discoveries of those specifications of neural computations that we now lack.

5. REDUCTION OF COGNITION TO THE SUBCONCEPTUAL LEVEL

The previous section considered the relationship between the fundamental level of the subsymbolic paradigm – the subconceptual level – and the neural level. The remainder of this article will focus on relations between the subconceptual and conceptual levels; these have so far only been touched upon briefly (in (8c)). Before proceeding, however, it is worth summarizing the relationships between the levels, including those that will be discussed in the remainder of the article.

Imagine three physical systems: a brain that is executing some cognitive process, a massively parallel connectionist computer running a subsymbolic model of that process, and a von Neumann computer running a symbolic model of the same process. The cognitive process may involve conscious rule application, intuition, or a combination of the two. According to the subsymbolic paradigm, here are the relationships:

- (13) a. Describing the brain at the neural level gives a neural model.
- b. Describing the brain approximately, at a higher level – the subconceptual level – yields, to a good approximation, the model running on the connectionist computer, when it too is described at the subconceptual level. (At this point, this is a goal for future research. It could turn out that the degree of approximation here is only rough; this would still be consistent with the subsymbolic paradigm.)
- c. We can try to describe the connectionist computer at a higher level – the conceptual level – by using the patterns of activity that have conceptual semantics. If the cognitive process being executed is conscious rule application, we will be able to carry out this conceptual-level analysis with reasonable precision, and will end up with a description that closely matches the symbolic computer program running on the von Neumann machine.
- d. If the process being executed is an intuitive process, we will be unable to carry out the conceptual-level description of the

connectionist machine precisely. Nonetheless, we will be able to produce various approximate conceptual-level descriptions that correspond to the symbolic computer program running on the von Neumann machine in various ways.

For a cognitive process involving both intuition and conscious rule application, (13c), and (13d) will each apply to certain aspects of the process.

The relationships (13a) and (13b) were discussed in the previous section. The relationship (13c) between a subsymbolic implementation of the conscious rule interpreter and a symbolic implementation is discussed in Section 6. The relations (13d) between subsymbolic and symbolic accounts of intuitive processing are considered in Section 9. These relations hinge on certain subsymbolic computational principles operative at the subconceptual level (13b); these are briefly discussed in Section 8. These principles are of a new kind for cognitive science, giving rise to the foundational considerations taken up in Section 7.

The relationships in (13) can be more clearly understood by reintroducing the concept of "virtual machine." If we take one of the three physical systems and describe its processing at a certain level of analysis, we get a virtual machine that I will denote "system_{level}". Then (13) can be written:

- (14) a. brain_{neural} = neural model
- b. brain_{subconceptual} ≈ connectionist_{subconceptual}
- c. connectionist_{conceptual} ≈ von Neumann_{conceptual}
(conscious rule application)
- d. connectionist_{conceptual} ~ von Neumann_{conceptual} (intuition)

Here, the symbol \approx means "equals to a good approximation" and \sim means "equals to a crude approximation." The two nearly equal virtual machines in (14c) both describe what I have been calling the "conscious rule interpreter." The two roughly similar virtual machines in (14d) provide the two paradigms' descriptions of the intuitive processor at the conceptual level.

Table II indicates these relationships and also the degree of exactness to which each system can be described at each level – the degree of precision to which each virtual machine is defined. The levels included in Table II are those relevant to predicting high-level behavior. Of course each system can also be described at lower levels, all the way down to elementary particles. However, levels below an exactly describable level can be ignored from the point of view of predicting high-level behavior, since it is possible (in principle) to do the prediction at the highest level that can be exactly

described (it is presumably much harder to do the same at lower levels). This is why in the symbolic paradigm any descriptions below the conceptual level are not viewed as significant. For modeling high-level behavior, how the symbol manipulation happens to be implemented can be ignored – it is not a relevant part of the cognitive model. In a subsymbolic model, exact behavioral prediction must be performed at the subconceptual level, but how the units happen to be implemented is not relevant.

TABLE II
Three cognitive systems and three levels of description

Level	(process)	Cognitive system		
		Brain	Subsymbolic	Symbolic
Conceptual	(intuition)	?	rough approximation	~ exact
	(conscious rule application)	?	good approximation	≈ exact
Subconceptual		good approximation	≈ exact	
Neural		exact		

The relation between the conceptual level and lower levels is fundamentally different in the subsymbolic and symbolic paradigms. This leads to important differences in the kind of explanations the paradigms offer of conceptual-level behavior, and the kind of reduction used in these explanations. A symbolic model is a *system* of interacting processes, all with the same conceptual-level semantics as the task behavior being explained. Adopting the terminology of Haugeland (1978), this *systematic explanation* relies on a *systematic reduction* of the behavior that involves no shift of semantic domain or *dimension*. Thus a game-playing program is composed of subprograms that generate possible moves, evaluate them, and so on. In the symbolic paradigm, these systematic reductions play the major role in explanation. The lowest-level processes in the systematic reduction, still with the original semantics of the task domain, are then themselves reduced by *intentional instantiation*: they are implemented exactly by other processes with different semantics but the same form. Thus a move-generation subprogram with game semantics is instantiated in a system of programs with list-manipulating semantics. This intentional instantiation typically plays a minor role in the overall explanation, if indeed it is regarded as a cognitively relevant part of the model at all.

Thus cognitive explanations in the symbolic paradigm rely primarily on

reductions involving no dimensional shift. This feature is not shared by the subsymbolic paradigm, where accurate explanations of intuitive behavior require descending to the subconceptual level. The elements in this explanation, the units, do *not* have the semantics of the original behavior: that is the content of the subconceptual unit hypothesis, (8b). In other words:

- (15) Unlike symbolic explanations, subsymbolic explanations rely crucially on a semantic ("dimensional") shift that accompanies the shift from the conceptual to the subconceptual levels.

The overall dispositions of cognitive systems are explained in the subsymbolic paradigm as approximate higher-level regularities that emerge from quantitative laws operating at a more fundamental level with different semantics. This is the kind of reduction familiar in natural science, exemplified by the explanation of the laws of thermodynamics through a reduction to mechanics that involves shifting the dimension from thermal semantics to molecular semantics. (Section 9 discusses some explicit subsymbolic reductions of symbolic explanatory constructs.)

Indeed the subsymbolic paradigm repeals the other features that Haugeland identified as newly introduced into scientific explanation by the symbolic paradigm. The inputs and outputs of the system are not quasilinguistic representations but good old-fashioned numerical vectors. These inputs and outputs have semantic interpretations, but these are not constructed recursively from interpretation of embedded constituents. The fundamental laws are good old-fashioned numerical equations.

Haugeland went to considerable effort to legitimize the form of explanation and reduction used in the symbolic paradigm. The explanations and reductions of the subsymbolic paradigm, by contrast, are of a type well-established in natural science.

In summary, let me emphasize that in the subsymbolic paradigm, the conceptual and subconceptual levels are not related as the levels of a von Neumann computer (high-level-language program, compiled low-level program, etc.). The relationship between subsymbolic and symbolic models is more like that between quantum and classical mechanics. Subsymbolic models accurately describe the microstructure of cognition, whereas symbolic models provide an approximate description of the macrostructure. An important job of subsymbolic theory is to delineate the situations and the respects in which the symbolic approximation is valid, and to explain why.

6. CONSCIOUS RULE APPLICATION IN THE SUBSYMBOLIC PARADIGM

In the symbolic paradigm, both conscious rule application and intuition are described at the conceptual level; that is, conscious and unconscious rule interpretation, respectively. In the subsymbolic paradigm, conscious rule application can be formalized in the conceptual level but intuition must be formalized at the subconceptual level. This suggests that a subsymbolic model of a cognitive process that involves both intuition and conscious rule interpretation would consist of two components using quite different formalisms. While this hybrid formalism might have considerable practical value, there are some theoretical problems with it. How would the two formalisms communicate? How would the hybrid system evolve with experience, reflecting the development of intuition and the subsequent remission of conscious rule application? How would the hybrid system elucidate the fallibility of actual human rule application (e.g. logic)? How would the hybrid system get us closer to understanding how conscious rule application is achieved neurally?

All these problems can be addressed by adopting a unified subconceptual-level analysis of both intuition and conscious rule interpretation. The virtual machine that is the conscious rule interpreter is to be implemented in a lower-level virtual machine: the same connectionist dynamical system that models the intuitive processor. How this can, in principle, be achieved is the subject of this section. The relative advantages and disadvantages of implementing the rule interpreter in a connectionist dynamical system, rather than a von Neumann machine, will also be considered.

Section 2.1 described the power of natural language for the propagation of cultural knowledge and the instruction of novices. Someone who has mastered a natural language has a powerful trick available for performing in domains where experience has been insufficient for the development of intuition: Verbally expressed rules, whether resident in memory or on paper, can be used to direct a step-by-step course to an answer. Once subsymbolic models have achieved a sufficient subset of the power to process natural language, they will be able to exploit the same trick. A subsymbolic system with natural language competence will be able to encode linguistic expressions as patterns of activity; like all other patterns of activity, these can be stored in connectionist memories using standard procedures. If the linguistic expressions stored in memory happen to be rules, the subsymbolic system can use them to solve problems sequentially in the following way. Suppose, for concreteness, that the rules stored in memory are production rules of the

form “if *condition* holds, then do *action*.” If the system finds itself in a particular situation where *condition* holds, then the stored production can be retrieved from the connectionist memory via the characteristic *content-addressability* of these memories: of the activity pattern representing the entire production, the subpart that pertains to *condition* is present, and this then leads to the reinstatement in the memory of the entire pattern representing the production. The competence of the subsymbolic system to process natural language must include the ability to take the portion of the reinstated pattern that encodes the verbal description of *action*, and actually execute the action it describes; that is, the subsymbolic system must be able to *interpret*, in the computational sense of the term, the memorized description of *action*. The result is a subsymbolic implementation of a production system, built purely out of subsymbolic natural language processing mechanisms. A connectionist account of natural language processes must eventually be developed as part of the subsymbolic paradigm, because natural language processes of fluent speakers are intuitive and thus, according to the subsymbolic hypothesis (8), must be modeled at the subconceptual level using subsymbolic computation.

In summary:

- (16) The competence to represent and process linguistic structures in a native language is a competence of the human intuitive processor; the subsymbolic paradigm assumes that this competence can be modeled in a subconceptual connectionist dynamical system. By combining such linguistic competence with the memory capabilities of connectionist systems, sequential rule interpretation can be implemented.

Now note that our subsymbolic system can use its stored rules to perform the task. The standard learning procedures of connectionist models now turn this experience of performing the task into a set of weights for going from inputs to outputs. Eventually, after enough experience, the task can be performed directly by these weights. The input activity generates the output activity so quickly that before the relatively slow rule-interpretation process has a chance to reinstantiate the first rule in memory and interpret it, the task is done. With intermediate amounts of experience, some of the weights are well enough in place to prevent some of the rules from having the chance to instantiate, while others are not, enabling other rules to be retrieved and interpreted.

6.1. Rule Interpretation, Consciousness, and Seriality

What about the conscious aspect of rule interpretation? Since consciousness seems to be a quite high-level description of mental activity, it is reasonable to suspect that it reflects the very coarse structure of the cognitive dynamical system. This suggests the following hypothesis:

- (17) The contents of consciousness reflect only the large-scale structure of activity patterns: subpatterns of activity that are extended over spatially large regions of the network and that are stable for relatively long periods of time.

(See Rumelhart, Smolensky, McClelland and Hinton 1986. Note that (17) hypothesizes a *necessary* – not a *sufficient* – condition for an aspect of the subsymbolic state to be relevant to the conscious state.) The spatial aspect of this hypothesis has already played a major role in this article – it is in fact a restatement of the subconceptual unit hypothesis, (8b): Concepts that are consciously accessible correspond to patterns over large numbers of units. It is the temporal aspect of hypothesis (17) that is relevant here. The rule interpretation process requires that the retrieved linguistically coded rule be maintained in memory while it is being interpreted. Thus the pattern of activity representing the rule must be stable for a relatively long time. In contrast, after connections have been developed to perform the task directly, there is no correspondingly stable pattern formed during the performance of the task. Thus the loss of conscious phenomenology with expertise can be understood naturally.

On this account, the sequentiality of the rule interpretation process is not built into the architecture; rather, it is linked to our ability to follow only one verbal instruction at a time. Connectionist memories have the ability to retrieve a single stored item, and here this ability is called upon so that the linguistic interpreter is not required to interpret multiple instructions simultaneously.

It is interesting to note that the preceding analysis also applies to nonlinguistic rules: Any notational system that can be appropriately interpreted will do. For example, another type of rule might be a short series of musical pitches; a memorized collection of such rules would allow a musician to play a tune by conscious rule interpretation. With practice, the need for conscious control goes away. Since pianists learn to interpret several notes simultaneously, the present account suggests that a pianist might be able to apply more than one musical rule at a time; if the pianist's memory for these rules

can simultaneously recall more than one, it would be possible to generate multiple musical lines simultaneously using conscious rule interpretation. A symbolic account of such a process would involve something like a production system capable of firing multiple productions simultaneously.

Finally, it should be noted that even if the memorized rules are assumed to be linguistically coded, the preceding analysis is uncommitted about the form the encoded rules take in memory: phonological, orthographic, semantic, or whatever.

6.2. Symbolic Versus Subsymbolic Implementation of Rule Interpretation

The (approximate) implementation of the conscious rule interpreter in a subsymbolic system has both advantages and disadvantages relative to an (exact) implementation in a von Neumann machine.

The main disadvantage is that subconceptual representation and interpretation of linguistic instructions is very difficult and we are not actually able to do it now. Most existing subsymbolic systems simply don't use rule interpretation.⁸ Thus they miss out on all the advantages listed in (2). They can't take advantage of rules to check the results produced by the intuitive processor. They can't bootstrap their way into a new domain using rules to generate their own experience: they must have a teacher generate it for them.⁹

There are several advantages of a subconceptually implemented rule interpreter. The intuitive processor and rule interpreter are highly integrated, with broad-band communication between them. Understanding how this communication works should allow the design of efficient hybrid symbolic/subsymbolic systems with effective communication between the processors. A principled basis is provided for studying how rule-based knowledge leads to intuitive knowledge. Perhaps most interesting, in a subsymbolic rule interpreter, the process of rule selection is intuitive! Which rule is reinstated in memory at a given time is the result of the associative retrieval process, which has many nice properties. The best match to the productions' conditions is quickly computed, and even if no match is very good, a rule can be retrieved. The selection process can be quite context-sensitive.

An integrated subsymbolic rule interpreter/intuitive processor in principle offers the advantages of both kinds of processing. Imagine such a system creating a mathematical proof. The intuitive processor would generate goals and steps, and the rule interpreter would verify their validity. The serial

search through the space of possible steps, which is necessary in a purely symbolic approach, is replaced by the intuitive generation of possibilities. Yet the precise adherence to strict inference rules that is demanded by the task can be enforced by the rule interpreter; the creativity of intuition can be exploited while its unreliability can be controlled.

6.3. Two Kinds of Knowledge – One Knowledge Medium

Most existing subsymbolic systems perform tasks without serial rule interpretation: Patterns of activity representing inputs are directly transformed (possibly through multiple layers of units) into patterns of activity representing outputs. The connections that mediate this transformation represent a form of task knowledge that can be applied with massive parallelism: I will call it *P-knowledge*. For example, the P-knowledge in a native speaker presumably encodes lexical, morphological, syntactic, semantic, and pragmatic constraints in such a form that all these constraints can be satisfied in parallel during comprehension and generation.

The connectionist implementation of sequential rule interpretation described above displays a second form that knowledge can take in a subsymbolic system. The stored activity patterns that represent rules also constitute task knowledge: Call it *S-knowledge*. Like P-knowledge, S-knowledge is embedded in connections: the connections that enable part of a rule to reinstantiate the entire rule. Unlike P-knowledge, S-knowledge cannot be used with massive parallelism. For example, a novice speaker of some language cannot satisfy the constraints contained in two memorized rules simultaneously; they must be serially reinstated as patterns of activity and separately interpreted. Of course, the connections responsible for reinstating these memories operate in parallel, and indeed these connections contain within them the potential to reinstantiate either of the two memorized rules. But these connections are so arranged that only one rule at a time can be reinstated. The retrieval of each rule is a parallel process, but the satisfaction of the constraints contained within the two rules is a serial process. After considerable experience, P-knowledge is created: connections that can simultaneously satisfy the constraints represented by the two rules.

P-knowledge is considerably more difficult to create than S-knowledge. To encode a constraint in connections so that it can be satisfied in parallel with thousands of others is not an easy task. Such an encoding can only be learned through considerable experience in which that constraint has appeared in many different contexts, so that the connections enforcing the constraint can

be tuned to operate in parallel with those enforcing a wide variety of other constraints. S-knowledge can be acquired (once the linguistic skills on which it depends have been encoded into P-knowledge, of course) much more rapidly. For example, simply reciting a verbal rule over and over will usually suffice to store it in memory, at least temporarily.

That P-knowledge is so highly context-dependent while the rules of S-knowledge are essentially context-independent is an important computational fact underlying many of the psychological explanations offered by subsymbolic models. Consider, for example, Rumelhart and McClelland's (1986) model of the U-shaped curve for past-tense production in children. The phenomenon is striking: A child is observed using *goed* and *wented* when at a much younger age *went* was reliably used. This is surprising because we are prone to think that such linguistic abilities rest on knowledge that is encoded in some context-independent form such as "the past tense of *go* is *went*." Why should a child *lose* such a rule once acquired? A traditional answer invokes the acquisition of a different context-independent rule, such as "the past tense of *x* is *x + ed*" which, for one reason or another, takes precedence. The point here, however, is that there is nothing at all surprising about the phenomenon when the underlying knowledge is assumed to be context-dependent and not context-independent. The young child has a small vocabulary of largely irregular verbs. The connections that implement this P-knowledge are reliable in producing the large pattern of activity representing *went*, as well as those representing a small number of other past-tense forms. Informally we can say that the connections producing *went* do so in the context of the other vocabulary items that are also stored in the same connections. There is no guarantee that these connections will produce *went* in the context of a different vocabulary. As the child acquires additional vocabulary items, most of which are regular, the context radically changes. Connections that were, so to speak, perfectly adequate for creating *went* in the old context now have to work in a context where very strong connections are trying to create forms ending in *-ed*; the old connections are not up to the new task. Only through extensive experience trying to produce *went* in the new context of many regular verbs can the old connections be modified to work in the new context. In particular, strong new connections must be added that, when the input pattern encodes *go*, cancel the *-ed* in the output; these were not needed before.

These observations about context-dependence can also be framed in terms of inference. If we choose to regard the child as using knowledge to infer the correct answer *went*, then we can say that after the child has added more

knowledge (about new verbs), the ability to make the (correct) inference is lost. In this sense the child's inference process is nonmonotonic – perhaps this is why we find the phenomenon surprising. As will be discussed in Section 8, nonmonotonicity is a fundamental property of subsymbolic inference.

To summarize:

- (18) a. Knowledge in subsymbolic systems can take two forms, both resident in the connections.
- b. The knowledge used by the conscious rule interpreter lies in connections that reinstantiate patterns encoding rules; task constraints are coded in context-independent rules and satisfied serially.
- c. The knowledge used in intuitive processing lies in connections that constitute highly context-dependent encodings of task constraints that can be satisfied with massive parallelism.
- d. Learning such encodings requires much experience.

7. SUBSYMBOLIC DEFINITION OF COGNITIVE SYSTEMS AND SOME FOUNDATIONAL ISSUES

In order for the subconceptual level to be rightly viewed as a level for practicing cognitive science, it is necessary that the principles formulated at this level truly be principles of cognition. Since subsymbolic principles are neither conceptual-level nor neural-level principles, it is not immediately apparent what kind of cognitive principles they might be. The structure of subsymbolic models is that of a dynamical system; in what sense do these models embody principles of cognition rather than principles of physics?

What distinguishes those dynamical systems that are cognitive from those that are not? At this point the types of dynamical systems being studied in connectionist cognitive science lack anything that could justly be called an intentional psychology. In this section I wish to show that it is nonetheless possible to distinguish the sort of dynamical systems that have so far been the object of study in connectionist cognitive science from the dynamical systems that have traditionally been the subject matter of physics, and that the questions being studied are indeed questions of cognition.

A crucial property of cognitive systems broadly construed is that over a wide variety of environments they can maintain, at an adequately constant level, the degree to which a significant number of *goal conditions* are met.

Here I intend the teleological, rather than the intentional, sense of “goal.” A river, for example, is a complex dynamical system that responds sensitively to its environment – but about the only condition that it can satisfy over a large range of environments is going downhill. A cockroach manages, over an annoyingly extensive range of environments, to maintain its nutritive intake, its reproductive demands, its oxygen intake, even its probability of getting smashed, all within a relatively narrow band. The repertoire of conditions that people can keep satisfied, and the range of environments under which this relative constancy can be maintained, provides a measure worthy of the human cognitive capacity.

- (19) Cognitive system: A necessary condition for a dynamical system to be *cognitive* is that, under a wide variety of environmental conditions, it maintains a large number of goal conditions. The greater the repertoire of goals and variety of tolerable environmental conditions, the greater the cognitive capacity of the system.

The issue of complexity is crucial here. A river (or a thermostat) only fails to be a cognitive dynamical system because it cannot satisfy a *large* range of goals under *wide* range of conditions.¹⁰ Complexity is largely what distinguishes the dynamical systems studied in the subsymbolic paradigm from those traditionally studied in physics. Connectionist dynamical systems have great complexity: The information content in their weights is very high. Studying the extent to which a connectionist dynamical system can achieve complex goals in complex environments requires grappling with complexity in dynamical systems in a way that is traditionally avoided in physics. In cognitive modeling, many of the basic questions concern the detailed dynamics of a distinct pattern of activation in a system with a particular initial state and a particular set of interaction strengths that are highly nonhomogeneous. This is like asking a physicist: “Suppose we have a gas with 10,000 particles with the following 10,000 different masses and the following 500,000 different forces between them. Suppose we start them at rest in the following 10,000 positions. What are the trajectories of the following 20 particles?” This is indeed a question about a dynamical system, and is, in a sense, a question of physics. It is this kind of question, however, that is avoided at all costs in physics. The physicist we consulted is likely to compute the mean collision times for the particles assuming equal masses, random starting positions, and uniformly random interactions, and say “if that isn’t good enough, then take your question to a computer.”¹¹

Nonetheless, physics has valuable concepts and techniques to contribute to the study of connectionist dynamical systems. Insights from physics have already proved important in various ways in the subsymbolic paradigm (Hinton and Sejnowski 1983a; Sejnowski 1976; Smolensky 1983).

Various subsymbolic models have addressed various goals and environments. A very general goal that is of particular importance is:

- (20) *The prediction goal:* Given some partial information about the environmental state, correctly infer missing information.

What is maintained here is the degree of match between predicted values and the actual values for the unknowns. Maintenance of this match over the wide range of conditions found in a complex environment is a difficult task. Special cases of this task include predicting the depth of an object from retinal images, the future location of a moving object, the change in certain aspects of an electric circuit given the changes in other aspects, or the propositions implied by a text. The prediction goal is obviously an important one, because it can serve so many other goals: Accurate prediction of the effects of actions allows the selection of those leading to desired effects.

A closely related goal is:

- (21) *The prediction-from-examples goal:* Given more and more examples of states from an environment, achieve the prediction goal with increasing accuracy in that environment.

For the prediction goal we ask: What inference procedures and knowledge about an environment must a dynamical system possess to be able to predict that environment? For the prediction-from-examples goal we go further and ask: What learning procedures must a dynamical system possess to be able to acquire the necessary knowledge about an environment from examples?

The goals of prediction and prediction-from-examples are the subject of many principles of the subsymbolic paradigm. These are indeed cognitive principles. They will be taken up in the next section; first, however, I would like to consider some implications of this characterization of a cognitive system for certain foundational issues: semantics, rationality, and the constituent structure of mental states. It would be absurd to suggest that the following few paragraphs constitute definitive treatments of these issues; the intent is rather to indicate specific points where subsymbolic research touches on these issues and to sow seeds for further analysis.

7.1. Semantics and Rationality in the Subsymbolic Paradigm

The subsymbolic characterization of a cognitive system (19) intrinsically binds cognitive systems both to states of the environment and to goal conditions. It therefore has implications for the question: How do states of a subsymbolic system get their meanings and truth conditions? A starting point for an answer is suggested in the following hypothesis:

- (22) Subsymbolic semantics: A cognitive system adopts various internal states in various environmental conditions. To the extent that the cognitive system meets its goal conditions in various environmental conditions, its internal states are *veridical representations* of the corresponding environmental states, with respect to the given goal conditions.

For the prediction goal, for example, a state of the subsymbolic system is a veridical representation of the current environmental state to the extent that it leads to correct predictions.

According to hypothesis (22), it is not possible to localize a failure of veridical representation. Any particular state is part of a large causal system of states, and failures of the system to meet goal conditions cannot in general be localized to any particular state or state component.¹² In subsymbolic systems, this *assignment of blame problem* (Minsky 1963) is a difficult one, and it makes programming subsymbolic models by hand very tricky. Solving the assignment of blame problem is one of the central accomplishments of the automatic network programming procedures: the learning procedures of the subsymbolic paradigm.

The characterization (19) of cognitive systems relates to rationality as well. How can one build a rational machine? How can internal processes (e.g., inference) be guaranteed to preserve veridical semantic relationships (e.g., be truth preserving)? These questions now become: How can the connection strengths be set so that the subsymbolic system will meet its goal conditions? Again, this is a question answered by the scientific discoveries of the subsymbolic paradigm: particular procedures for programming machines to meet certain goals – especially learning procedures to meet adaptation goals such as prediction-from-examples.

Let me compare this subsymbolic approach to veridicality with a symbolic approach to truth preservation offered by Fodor (1975; 1987). In the context of model-theoretic semantics for a set of symbolic formulae, proof theory provides a set of symbol manipulations (rules of inference) guaranteed to

preserve truth conditions. Thus if an agent possesses knowledge in the symbolic form $p \rightarrow q$ and additional knowledge p , then by syntactic operations the agent can produce q ; proof theory guarantees that the truth conditions of the agent's knowledge (or beliefs) has not changed.

There are fairly direct subsymbolic counterparts to this proof theoretic account. The role of logical inference is played by statistical inference. By explicitly formalizing tasks like prediction as statistical inference tasks, it is possible to prove for appropriate systems that subsymbolic computation is valid in a sense directly comparable to symbolic proof. Further discussion of this point, which will appear in Section 9.1, must await further examination of the computational framework of the subsymbolic paradigm, which is the subject of Section 8.

Note that the proof theoretic account explains the tautological inference of q from p and $p \rightarrow q$, but it leaves to an independent module an account of how the agent acquired the knowledge $p \rightarrow q$ that licenses the inference from p to q . In the subsymbolic account, the veridicality problem is tied inextricably to the environment in which the agent is trying to satisfy the goal conditions – subsymbolic semantics is *intrinsically* situated. The subsymbolic analysis of veridicality involves the following basic questions: How can a cognitive system be put in a novel environment and learn to create veridical internal representations that allow valid inferences about that environment so that goal conditions can be satisfied? How can it pick up information from its environment? These are exactly the questions addressed by subsymbolic learning procedures.

Note that in the subsymbolic case, the internal processing mechanisms (which can appropriately be called inference procedures) do not, of course, directly depend causally on the environmental state that may be internally represented or on the veridicality of that representation. In that sense, they are just as formal as syntactic symbol manipulations. The fact that a subsymbolic system can generate veridical representations of the environment (e.g., make valid predictions) is a result of extracting information from the environment and internally coding it in its weights through a learning procedure.

7.2. *Constituent Structure of Mental States*

Fodor and Pylyshyn have argued (e.g., Fodor 1975; Pylyshyn 1984) that mental states must have constituent structure, and they have used this argument against the connectionist approach (Fodor and Pylyshyn 1988). Their argument applies, however, only to ultra-local connectionist models

(Ballard and Hayes 1984); it is quite inapplicable to the distributed connectionist systems considered here. A mental state in a subsymbolic system is a pattern of activity with a constituent structure that can be analyzed at both the conceptual and the subconceptual levels. In this section I offer a few general observations on this issue; the connectionist representation of complex structures is an active area of research (Smolensky 1987; Touretzky 1986), and many difficult problems remain to be solved (for further discussion see Smolensky 1988).

At the conceptual level, a connectionist mental state contains constituent subpatterns that have conceptual interpretations. Pylyshyn, in a debate over the connectionist approach at the 1984 meeting of the Cognitive Science Society, suggested how to extract these conceptual constituents with the following example: The connectionist representation of *coffee* is the representation of *cup with coffee* minus the representation of *cup without coffee*. To carry out this suggestion, imagine a crude but adequate kind of distributed semantic representation, in which the interpretation of *cup with coffee* involves the activity of network units representing features like brown liquid with flat top surface, brown liquid with curved sides and bottom surface, brown liquid contacting porcelain, hot liquid, upright container with a handle, burnt odor, and so forth. We should really use subconceptual features, but even these features are sufficiently low-level to make the point. Following Pylyshyn, we take this representation of the interpretation of *cup with coffee* and subtract from it the representation of the interpretation of *cup without coffee*, leaving the representation of *coffee*. What remains, in fact, is a pattern of activity with active features such as brown liquid with flat top surface, brown liquid with curved sides and bottom surface, brown liquid contacting porcelain, hot liquid, and burnt odor. This represents *coffee*, in some sense – but *coffee in the context of cup*.

In using Pylyshyn's procedure for determining the connectionist representation of *coffee*, there is nothing sacred about starting with *cup with coffee*: why not start with *can with coffee*, *tree with coffee*, or *man with coffee*, and subtract the corresponding representation of *X without coffee*? Thinking back to the distributed featural representation, it is clear that each of these procedures produces quite a different result for "the" connectionist representation of *coffee*. The pattern representing *coffee* in the context of *cup* is quite different from the pattern representing *coffee* in the context of *can*, *tree*, or *man*.

The pattern representing *cup with coffee* can be decomposed into conceptual-level constituents, one for *coffee* and another for *cup*. This

decomposition differs in two significant ways from the decomposition of the symbolic expression *cup with coffee*, into the three constituents, *coffee*, *cup*, and *with*. First, the decomposition is quite approximate. The pattern of features representing *cup with coffee* may well, as in the imagined case above, possess a subpattern that can be identified with *coffee*, as well as a subpattern that can be identified with *cup*; but these subpatterns will in general not be defined precisely and there will typically remain features that can be identified only with the interaction of the two (as in brown liquid contacting porcelain). Second, whatever the subpattern identified with *coffee*, unlike the symbol *coffee*, it is a context-dependent constituent, one whose internal structure is heavily influenced by the structure of which it is a part.

These constituent subpatterns representing *coffee* in varying contexts are activity vectors that are not identical, but possess a rich structure of commonalities and differences (a family resemblance, one might say). The commonalities are directly responsible for the common processing implications of the interpretations of these various phrases, so the approximate equivalence of the *coffee* vectors across contexts plays a functional role in subsymbolic processing that is quite close to the role played by the exact equivalence of the *coffee* tokens across different contexts in a symbolic processing system.

The conceptual-level constituents of mental states are activity vectors, which themselves have constituent structure at the subconceptual level: the individual units' activities. To summarize the relationship between these notions of constituent structure in the symbolic and subsymbolic paradigms, let's call each *coffee* vector the (connectionist) symbol for coffee in the given context. Then we can say that the context alters the internal structure of the symbol; the activities of the subconceptual units that comprise the symbol – its subsymbols – change across contexts. In the symbolic paradigm, a symbol is effectively contextualized by surrounding it with other symbols in some larger structure. In other words:

- (23) Symbols and context dependence: In the symbolic paradigm, the context of a symbol is manifest around it and consists of other symbols; in the subsymbolic paradigm, the context of a symbol is manifest inside it and consists of subsymbols.

(Compare Hofstadter 1979; 1985.)

8. COMPUTATION AT THE SUBCONCEPTUAL LEVEL

Hypothesis (8a) offers a brief characterization of the connectionist architecture assumed at the subconceptual level by the subsymbolic paradigm. It is time to bring out the computational principles implicit in that hypothesis.

8.1. *Continuity*

According to (8a), a connectionist dynamical system has a continuous space of states and changes state continuously in time. I take time in this section to motivate at some length this assumption of continuity, because it plays a central role in the characterization of subsymbolic computation and because readers familiar with the literature on connectionist models will no doubt require that I reconcile the continuity assumption with some salient candidate counterexamples.

Within the symbolic paradigm, the simplest, most straightforward formalizations of a number of cognitive processes have quite discrete characters:

- (24) a. Discrete memory locations, in which items are stored without mutual interaction.
- b. Discrete memory storage and retrieval operations, in which an entire item is stored or retrieved in a single, atomic (primitive) operation.
- c. Discrete learning operations, in which new rules become available for use in an all-or-none fashion.
- d. Discrete inference operations, in which conclusions become available for use in an all-or-none fashion.
- e. Discrete categories, to which items either belong or do not belong.
- f. Discrete production rules, with conditions that are either satisfied or not satisfied, and actions that either execute or do not execute.

These discrete features come “for free” in the symbolic paradigm: Of course, any one of them can be softened but only by explicitly building in machinery to do so.

Obviously (24) is a pretty crude characterization of cognitive behavior. Cognition seems to be a richly interwoven fabric of graded, continuous processes and discrete, all-or-none processes. One way to model this interplay is to posit separate discrete and continuous processors in interaction. Some theoretical problems with this move were mentioned in Section 6,

where a unified formalism was advocated. It is difficult to introduce a hard separation between the soft and the hard components of processing. An alternative is to adopt a fundamentally symbolic approach, but to soften various forms of discreteness by hand. For example, the degree of match to conditions of production rules can be given numerical values, productions can be given strengths, interactions between separately stored memory items can be put in by hand, and so on (Anderson 1983).

The subsymbolic paradigm offers another alternative. All the discrete features of (24) are neatly swept aside in one stroke by adopting a continuous framework that applies at the subconceptual level. Then, when the continuous system is analyzed at the higher, conceptual level, various aspects of discreteness emerge naturally and inevitably, without explicit machinery having been devised to create this discreteness. These aspects of "hardness" are intrinsically embedded in a fundamentally "soft" system. The dilemma of accounting for both the hard and soft aspects of cognition is solved by using the passage from a lower level of analysis to a higher level to introduce natural changes in the character of the system: The emergent properties can have a different nature from the fundamental properties. This is the story to be fleshed out in the remainder of the paper. It rests on the fundamental continuity of subsymbolic computation, which is further motivated in the remainder of this section (for further discussion see Smolensky 1988).

It may appear that the continuous nature of subsymbolic systems is contradicted by the fact that it is easy to find in the connectionist literature models that are quite within the spirit of the subsymbolic paradigm, but which have neither continuous state spaces nor continuous dynamics. For example, models having units with binary values that jump discretely on the ticks of a discrete clock (the Boltzmann machine, Ackley et al. 1985; Hinton and Sejnowski 1983a; harmony theory, Smolensky 1983; 1986a). I will now argue that these models should be viewed as discrete simulations of an underlying continuous model, considering first discretization of time and then discretization of the units' values.

Dynamical systems evolving in continuous time are almost always simulated on digital computers by discretizing time. Since subsymbolic models have almost always been simulated on digital computers, it is no surprise that they too have been simulated by discretizing time. The equations defining the dynamics of the models can be understood more easily by most cognitive scientists if the differential equations of the underlying continuous dynamical system are avoided in favor of the discrete-time approximations that actually get simulated.

When subsymbolic models use binary-valued units, these values are best viewed not as symbols like *T* and *NIL* that are used for conditional branching tests, but as numbers (not numerals!) like 1 and 0 that are used for numerical operations (e.g., multiplication by weights, summation, exponentiation). These models are formulated in such a way that they are perfectly well-defined for continuous values of the units. Discrete numerical unit values are no more than a simplification that is sometimes convenient.¹³

As historical evidence that underlying subsymbolic models are continuous systems, it is interesting to note that when the theoretical conditions that license the discrete approximation have changed, the models have reverted to continuous values. In the harmony/energy optimal model, when the jumpy stochastic search was replaced by a smooth deterministic one (Rumelhart, Smolensky, McClelland and Hinton 1986), the units were changed to continuous ones.¹⁴

A second, quite dramatic, piece of historical evidence is a case where switching from discrete to continuous units made possible a revolution in subsymbolic learning theory. In their classic book, *Perceptrons*, Minsky and Papert (1969) exploited primarily discrete mathematical methods that were compatible with the choice of binary units. They were incapable of analyzing any but the simplest learning networks. By changing the discrete threshold function of perceptrons to a smooth, differentiable curve, and thereby defining continuous-valued units, Rumelhart, Hinton, and Williams (1986) were able to apply continuous analytic methods to more complex learning networks. The result was a major advance in the power of subsymbolic learning.

A third historical example of the power of a continuous conception of subsymbolic computation relates to the connectionist generation of sequences. Traditionally this task has been viewed as making a connectionist system jump discretely between states to generate an arbitrary discrete sequence of actions A_1, A_2, \dots This view of the task reduces the connectionist system to a finite state machine that can offer little new to the analysis of sequential behavior. Recently Jordan (1986) has shown how a subsymbolic approach can give “for free” co-articulation effects where the manner in which actions are executed is influenced by future actions. Such effects are just what should come automatically from implementing serial behavior in a fundamentally parallel machine. Jordan’s trick is to view the connectionist system as evolving continuously in time, with the task being the generation of a continuous trajectory through state space, a trajectory that meets as boundary conditions certain constraints, for example, that the discrete times

1, 2, ... the system state must be in regions corresponding to the actions A_1, A_2, \dots .

The final point is a foundational one. The theory of discrete computation is quite well understood. If there is any new theory of computation implicit in the subsymbolic approach, it is likely to be a result of a fundamentally different, continuous formulation of computation. It therefore seems fruitful, in order to maximize the opportunity for the subsymbolic paradigm to contribute new computational insights, to hypothesize that subsymbolic computation is fundamentally continuous.

It must be emphasized that the discrete/continuous distinction cannot be understood completely by looking at simulations. Discrete and continuous machines can of course simulate each other. The claim here is that the most analytically powerful descriptions of subsymbolic models are continuous ones, whereas those of symbolic models are not continuous.

This has profound significance because it means that many of the concepts used to understand cognition in the subsymbolic paradigm come from the category of continuous mathematics, while those used in the symbolic paradigm come nearly exclusively from discrete mathematics. Concepts from physics, from the theory of dynamical systems, are at least as likely to be important as concepts from the theory of digital computation. And analog computers, both electronic and optical, provide natural implementation media for subsymbolic systems (Anderson 1986; Cohen 1986).

8.2. Subsymbolic Computation

An important illustration of the continuous/discrete mathematics contrast that distinguishes subsymbolic from symbolic computation is found in inference. A natural way to look at the knowledge stored in connections is to view each connection as a *soft constraint*. A positive (excitatory) connection from unit a to unit b represents a soft constraint to the effect that if a is active, then b should be too. A negative (inhibitory) connection represents the opposite constraint. The numerical magnitude of a connection represents the strength of the constraint.

Formalizing knowledge in soft constraints rather than hard rules has important consequences. Hard constraints have consequences singly; they are rules that can be applied separately and sequentially – the operation of each proceeding independently of whatever other rules may exist. But soft constraints have no implications singly; any one can be overridden by the others. It is only the entire set of soft constraints that has any implications.

Inference must be cooperative process, like the parallel relaxation processes typically found in subsymbolic systems. Furthermore, adding additional soft constraints can repeal conclusions that were formerly valid: Subsymbolic inference is fundamentally nonmonotonic.

One way of formalizing soft constraint satisfaction is in terms of statistical inference. In certain subsymbolic systems, the soft constraints can be identified as statistical parameters, and the activation passing procedures can be identified as statistical inference procedures (Geman and Geman 1984; Hinton and Sejnowski 1983b; Pearl 1985; Shastri 1985; Smolensky 1986a). This identification is usually rather complex and subtle: Unlike in classical “spreading activation” models and in many local connectionist models, the strength of the connection between two units is *not* determined solely by the correlation between their activity (or their “degree of association”). To implement subsymbolic statistical inference, the correct connection strength between two units will typically depend on all the other connection strengths. The subsymbolic learning procedures that sort out this interdependence through simple, strictly local, computations and ultimately assign the correct strength to each connection are performing no trivial task.

To sum up:

- (25) a. Knowledge in subsymbolic computation is formalized as a large set of soft constraints.
- b. Inference with soft constraints is fundamentally a parallel process.
- c. Inference with soft constraints is fundamentally nonmonotonic.
- d. Certain subsymbolic systems can be identified as using statistical inference.

9. CONCEPTUAL-LEVEL DESCRIPTIONS OF INTUITION

The previous section concerned computation in subsymbolic systems analyzed at the subconceptual-level, the level of units and connections. In this final section I consider analyses of subsymbolic computation at the higher, conceptual level. Section 6 discussed subsymbolic modeling of conscious rule interpretation; here I consider subsymbolic models of intuitive processes. I will elaborate the point foreshadowed in Section 5: Conceptual-level descriptions of aspects of subsymbolic models of intuitive processing roughly approximate symbolic accounts. The picture that emerges is of a symbiosis between the symbolic and subsymbolic paradigms: The symbolic paradigm

offers concepts for better understanding subsymbolic models, and those concepts are in turn illuminated with a fresh light by the subsymbolic paradigm.

9.1. The Best Fit Principle

The notion that each connection represents a soft constraint can be formulated at higher level:

- (26) The Best Fit Principle: Given an input, a subsymbolic system output a set of inferences that, as a whole, give a best fit to the input, in a statistical sense defined by the statistical knowledge stored in the system's connections.

In this vague form, this principle can be regarded as a desideratum of subsymbolic systems. Given the principle, formal embodiment in a class of connectionist dynamical systems was the goal of harmony theory (Riley and Smolensky 1984; Smolensky 1983; 1984a; 1984b; 1986a; 1986c).

To render the Best Fit Principle precise, it is necessary to provide precise definitions of “inferences,” “best fit,” and “statistical knowledge stored in the system’s connections.” This is done in harmony theory, where the central object is the harmony function H which measures, for any possible set of inferences, the goodness of fit to the input with respect to the soft constraints stored in the connection strengths. The set of inferences with the largest value of H , that is, highest harmony, is the best set of inferences, with respect to a well-defined statistical problem.

Harmony theory offers three things. It gives a mathematically precise characterization of the prediction-from-examples goal as a statistical inference problem. It tells how the prediction goal can be achieved using a network with a certain set of connections. Moreover, it gives a procedure by which the network can learn the correct connections with experience, thereby satisfying the prediction-from-examples goal.

The units in harmony networks are stochastic: The differential equations defining the system are stochastic. There is a system parameter called the *computational temperature* that governs the degree of randomness in the units’ behavior, it goes to zero as the computation proceeds. (The process is *simulated annealing*, like in the Boltzmann machine: Ackley *et al.* 1985; Hinton and Sejnowski 1983a, 1983b, 1986. See Rumelhart, McClelland and the PDP Research Group, 1986, p. 148, and Smolensky, 1986a, for the relations between harmony theory and the Boltzmann machine.)

9.2. *Productions, Sequential Processing, and Logical Inference*

A simple harmony model of expert intuition in qualitative physics was described by Riley and Smolensky (1984) and Smolensky (1986a, 1986c). The model answers questions such as, “What happens to the voltages in this circuit if I increase this resistor?” (The questions refer to a particular simple circuit; the model’s expertise is built in and not the result of learning.) This connectionist problem-solving system illustrates several points about the relations between subconceptual- and conceptual-level descriptions of subsymbolic computation.

Very briefly, the model looks like this. The state of the circuit is represented as a vector of activity over a set of network units we can call *circuit state feature units* – “feature units” for short. A subpart of this activity pattern represents whether the circuit’s *current* has gone up, down, or stayed the same; other subparts indicate what has happened to the *voltage drops*, and so on. Some of these subpatterns are fixed by the givens in the problem, and the remainder comprise the answer to be computed by the network. There is a second set of network units, called *knowledge atoms*, each of which corresponds to a subpattern of activity over feature units. The subpatterns of features encoded by knowledge atoms are those that can appear in representations of possible states of the circuit: They are subpatterns that are allowed by the laws of circuit physics. The system’s knowledge of Ohm’s Law, for example, is distributed over the many knowledge atoms whose subpatterns encode the legal feature combinations for current, voltage, and resistance. The connections in the network determine which feature subpattern corresponds to a given knowledge atom. The subpattern corresponding to knowledge atom α includes a positive (negative) value for a particular feature f if there is a positive (negative) connection between unit α and unit f ; the subpattern for α does not include f at all if there is no connection between α and f . All connections are two-way: Activity can propagate from feature units to knowledge atoms and vice versa. The soft constraints encoded by these connections, then, say that “if subpattern α is present, then feature f should be positive (negative), and vice versa.”

In the course of computing an answer to a question, the units in the network change their values hundreds of times. Each time a unit recomputes its value, we have a *microdecision*. As the network converges to a solution, it is possible to identify *macrodecisions*, each of which amounts to a commitment of part of the network to a portion of the solution. Each macrodecision is the result of many individual microdecisions. These macrodecisions are

approximately like the firing of production rules. In fact, these productions fire in essentially the same order as a symbolic forward-chaining inference system.¹⁵ One can measure the total amount of order in the system and see that there is a qualitative change in the system when the first microdecisions are made – the system changes from a disordered phase to an ordered one.

It is a corollary of the way this network embodies the problem domain constraints, and the general theorems of harmony theory, that the system, when given a well-posed problem and unlimited relaxation time, will always give the correct answer. So under that idealization, the *competence* of the system is described by *hard* constraints: Ohm's Law, Kirchoff's Law – the laws of simple circuits. It's as though the model had those laws written down inside it. However, as in all subsymbolic systems, the *performance* of the system is achieved by satisfying a large set of *soft* constraints. What this means is that if we depart from the ideal conditions under which hard constraints seem to be obeyed, the illusion that the system has hard constraints inside is quickly dispelled. The system can violate Ohm's Law if it has to, but if it needn't violate the law, it won't. Outside the idealized domain of well-posed problems and unlimited processing time, the system gives sensible performance. It isn't brittle the way that symbolic inference systems are. If the system is given an ill-posed problem, it satisfies as many constraints as possible. If it is given inconsistent information, it doesn't fall flat and deduce just anything at all. If it is given insufficient information, it doesn't sit there and deduce nothing at all. Given limited processing time, the performance degrades gracefully as well. All these features emerge "for free," as automatic consequences of performing inference in a subsymbolic system; no extra machinery is added on to handle the deviations from ideal circumstances.

Returning to a physics level analogy introduced in Section 5, we have a "quantum" system that appears to be "Newtonian" under the proper conditions. A system that has, at the microlevel, soft constraints satisfied in parallel, has at the macrolevel, under the right circumstances, to have hard constraints, satisfied serially. But it doesn't *really*, and if you go outside the Newtonian domain, you see that it's really been a quantum system all along.

This model exemplifies the competence/performance distinction as it appears in the subsymbolic paradigm. We have an inference system (albeit a very limited one) whose performance is completely characterizable at the subconceptual-level in terms of standard subsymbolic computation: massively parallel satisfaction of multiple soft constraints. The system is fundamentally soft. Just the same, the behavior of the system can be analyzed

at a higher level, and, under appropriate situations (well-posed problems), and under suitable processing idealizations (unlimited computation time), the competence of the system can be described in utterly different computational terms: The hard rules of the circuit domain. The competence theory is extremely important, but the performance theory uses radically different computational mechanisms.

The relation of the competence theory and the performance theory for this model can be viewed as follows. The behavior of the system is determined by its harmony function, which determines a surface or “landscape” of harmony values over the space of network states. In this landscape there are peaks where the harmony achieves its maximal value: These global maxima correspond to network states representing circuit conditions that satisfy all the laws of physics. The competence theory nicely describes the structure of this discrete constellation of global harmony maxima. But these maxima are a tiny subset of an extended harmony landscape in which they are embedded, and the network’s performance is a stochastic search over the harmony landscape for these peaks. The givens of a problem restrict the search to the portion of the space consistent with those givens. If the problem is well-posed, exactly one of the global harmony peaks will be accessible to the system. Given unlimited search time, the system will probably end up at this peak: This is the limit in which the performance theory is governed by the competence theory. As the search time is reduced, the probability of the system’s not ending up at the correct harmony peak increases. If insufficient information is given in the problem, multiple global harmony peaks will be accessible, and the system will converge to one of those peaks. If inconsistent information is given in the problem, none of the global harmony peaks will be accessible. But within the space of states accessible to the network there will be highest peaks of harmony – these peaks are not as high as the inaccessible global maxima; they correspond to network states representing circuit states that satisfy as many as possible of the circuit laws. As the network computes, it will converge toward these best-available peaks.

Subsymbolic computation is the evolution of a dynamical system. The input to the computation is a set of constraints on which states are accessible to the system (or, possibly, the state of the system at time zero). The dynamical system evolves in time under its defining differential equations; typically, it asymptotically approaches some equilibrium state – the output. The function relating the system’s input to its output is its competence theory. This function is extremely important to characterize. But it is quite different from the performance theory of the system, which is the differential equation

governing the system's moment-to-moment evolution. Relating the performance and competence of cognitive systems coincides with one of the principal tasks of dynamical systems theory: relating a system's local description (differential equations) to its global (asymptotic) behavior.

9.3. *Conceptual-Level Spreading Activation*

In Section 7.2 it was pointed out that states of a subsymbolic model can be approximately analyzed as superpositions of vectors with individual conceptual-level semantics. It is possible to approximately analyze connectionist dynamical systems at the conceptual level, using the mathematics of the superposition operation. If the connectionist system is purely linear (so that the activity of each unit is precisely a weighted sum of the activities of the units giving it input), it can easily be proved that the higher-level description obeys formal laws of just the same sort as the lower level: The computations at the subconceptual and conceptual levels are isomorphic. Linear connectionist systems are of limited computational power, however; most interesting connectionist systems are nonlinear. Nevertheless, most of these are in fact *quasilinear*: A unit's value is computed by taking the weighted sum of its inputs and passing this through a nonlinear function like a threshold or sigmoid. In quasi-linear systems, each unit combines its inputs linearly even though the effects of this combination on the unit's activity is nonlinear. Furthermore, the problem-specific knowledge in such systems is in the combination weights, that is, the linear part of the dynamical equations; and in learning systems, it is generally only these linear weights that adapt. For these reasons, even though the higher level is not isomorphic to the lower level in nonlinear systems, there are senses in which the higher level approximately obeys formal laws similar to the lower level. (For details, see Smolensky 1986b.)

The conclusion here is a rather different one from the preceding section, where we saw how there are senses in which higher-level characterizations of certain subsymbolic systems approximate productions, serial processing, and logical inference. Now what we see is that there are also senses in which the laws describing cognition at the conceptual level are activation-passing laws like those at the subconceptual-level but operating between units with individual conceptual semantics. Such semantic level descriptions of mental processing (which include *local* connectionist models; see note 3) have been of considerable value in cognitive science. We can now see how these "spreading activation" accounts of mental processing can fit into the

subsymbolic paradigm.

9.4. *Schemata*

The final conceptual-level notion I will consider is that of the *schema* (e.g., Rumelhart, 1980). This concept goes back at least to Kant (1787/1963) as a description of mental concepts and mental categories. Schemata appear in many AI systems in the forms of frames, scripts, or similar structures: They are prepackaged bundles of information that support inference in prototypical situations. [See also Arbib: 'Levels of Modeling of Mechanisms of Visually Guided Behavior', *BBS*, 10(3), 1987.]

Briefly, I will summarize work on schemata in connectionist systems reported in Rumelhart, Smolensky, McClelland and Hinton (1986) (see also Feldman 1981; Smolensky 1986a; 1986c). This work addressed the case of schemata for rooms. Subjects were asked to describe some imagined rooms using a set of 40 features like has-ceiling, has-window, contains-toilet, and so on. Statistics were computed on these data and were used to construct a network containing one node for each feature as well as connections computed from the statistical data.

The resulting network can perform inference of the same general kind as that carried out by symbolic systems with schemata for various types of rooms. The network is told that some room contains a ceiling and an oven; the question is, what else is likely to be in the room? The system settles down into a final state, and among the inferences contained in that final state are: the room contains a coffee cup but no fireplace, a coffee pot but no computer.

The inference process in this system is simply one of greedily maximizing harmony. [Cf. *BBS* multiple book review of Sperber and Wilson's *Relevance*, *BBS* 10(4).] To describe the inference of this system on a higher level, we can examine the global states of the system in terms of their harmony values. How internally consistent are the various states in the space? It's a 40-dimensional state space, but various 2-dimensional subspaces can be selected, and the harmony values there can be graphically displayed. The harmony landscape has various peaks; looking at the features of the state corresponding to one of the peaks, we find that it corresponds to a prototypical bathroom; others correspond to a prototypical office, and so on for all the kinds of rooms subjects were asked to describe. There are no *units* in this system for bathrooms or offices – there are just lower-level descriptors. The prototypical bathroom is a pattern of activation, and the system's recognition of its prototypicality is reflected in the harmony peak for that pattern. It is a

consistent, “harmonious” combination of features: better than neighboring points, such as one representing a bathroom without a bathtub, which has distinctly lower harmony.

During inference, this system climbs directly uphill on the harmony landscape. When the system state is in the vicinity of the harmony peak representing the prototypical bathroom, the inferences it makes are governed by the shape of the harmony landscape there. This shape is like a schema that governs inferences about bathrooms. (In fact, harmony theory was created to give a connectionist formalization of the notion of schema; see Smolensky, 1984b, 1986a, 1986c.) Looking closely at the harmony landscape, we can see that the terrain around the “bathroom” peak has many of the properties of a bathroom schema: variables and constants, default values, schemata embedded inside schemata, and even cross-variable dependencies, which are rather difficult to incorporate into symbolic formalizations of schemata. The system behaves as though it had schemata for bathrooms, offices, and so forth, even though they are not really there at the fundamental level: These schemata are strictly properties of a higher-level description. They are informal, approximate descriptions – one might even say they are merely metaphorical descriptions – of an inference process too subtle to admit such high-level descriptions with great precision. Even though these schemata may not be the sort of object on which to base a formal model, nonetheless they are useful descriptions that help us understand a rather complex inference system.

9.5. Summary

In this section the symbolic structures in the intuitive processor have been viewed as entities in high-level descriptions of cognitive dynamical systems. From this perspective, these structures assume rather different forms from those arising in the symbolic paradigm. To sum up:

- (27) a. Macroinference is not a process of firing a symbolic production but rather of qualitative state change in a dynamical system, such as a phase transition.
- b. Schemata are not large symbolic data structures but rather the potentially intricate shapes of harmony maxima.
- c. Categories (it turns out) are attractors in connectionist dynamical systems: states that “suck in” to a common place many nearby states, like peaks of harmony functions.

- d. Categorization is not the execution of a symbolic algorithm but rather the continuous evolution of the dynamical system – the evolution that drives states into the attractors that maximize harmony.
- e. Learning is not the construction and editing of formulae, but rather the gradual adjustment of connection strengths with experience, with the effect of slowly shifting harmony landscapes, adapting old and creating new concepts, categories, and schemata.

The heterogeneous assortment of high-level mental structures that have been embraced in this section suggests that the conceptual-level lacks formal unity. This is precisely what one expects of approximate higher-level descriptions, which, capturing different aspects of global properties, can have quite different characters. According to the subsymbolic paradigm, the unity underlying cognition is to be found not at the conceptual level, but rather at the subconceptual level, where relatively few principles in a single formal framework lead to a rich variety of global behaviors.

10. CONCLUSION

In this target article I have not argued for the validity of a connectionist approach to cognitive modeling, but rather for a particular view of the role a connectionist approach might play in cognitive science. An important question remains: Should the goal of connectionist research be to replace other methodologies in cognitive science? Here it is important to avoid the confusion discussed in Section 2.1. There I argued that for the purpose of science, it is sound to formalize knowledge in linguistically expressed laws and rules – but it does not follow therefore that knowledge in an individual's mind is best formalized by such rules. It is equally true that even if the knowledge in a native speaker's mind is well formalized by a huge mass of connection strengths, it does not follow that the science of language should be such a set of numbers. On the contrary, the argument of Section 2.1 implies that the science of language should be a set of linguistically expressed laws, to the maximal extent possible.

The view that the goal of connectionist research should be to replace other methodologies may represent a naive form of eliminative reductionism. Successful lower-level theories generally serve not to replace higher-level ones, but to enrich them, to explain their successes and failures, to fill in

where the higher-level theories are inadequate, and to unify disparate higher-level accounts. The goal of subsymbolic research should not be to replace symbolic cognitive science, but rather to explain the strengths and weaknesses of existing symbolic theory, to explain how symbolic computation can emerge out of nonsymbolic computation, to enrich conceptual-level research with new computational concepts and techniques that reflect an understanding of how conceptual-level theoretical constructs emerge from subconceptual computation, to provide a uniform subconceptual theory from which the multiplicity of conceptual theories can all be seen to emerge, to develop new empirical methodologies that reveal subconceptual regularities of cognitive behavior that are invisible at the conceptual level, and to provide new subconceptual-level cognitive principles that explain these regularities.

The rich behavior displayed by cognitive systems has the paradoxical character of appearing on the one hand tightly governed by complex systems of hard rules, and on the other to be awash with variance, deviation, exception, and a degree of flexibility and fluidity that has quite eluded our attempts at simulation. *Homo sapiens* is the rational animal, with a mental life ruled by the hard laws of logic – but real human behavior is riddled with strong nonrational tendencies that display a systematicity of their own. Human language is an intricate crystal defined by tight sets of intertwining constraints – but real linguistic behavior is remarkably robust under deviations from those constraints. This ancient paradox has produced a deep chasm in both the philosophy and the science of mind: on one side, those placing the essence of intelligent behavior in the hardness of mental competence; on the other, those placing it in the subtle softness of human performance.

The subsymbolic paradigm suggests a solution to this paradox. It provides a formal framework for studying how a cognitive system can possess knowledge which is fundamentally *soft*, but at the same time, under ideal circumstances, admit good higher-level descriptions that are undeniably *hard*. The passage from the lower, subconceptual level of analysis to the higher, conceptual level naturally and inevitably introduces changes in the character of the subsymbolic system: The computation that emerges at the higher level incorporates elements with a nature profoundly different from that of the fundamental computational processes.

To turn this story into a scientific reality, a multitude of serious conceptual and technical obstacles must be overcome. The story does, however, seem to merit serious consideration. It is to be hoped that the story's appeal will prove sufficient to sustain the intense effort that will be required to tackle the obstacles.

ACKNOWLEDGMENTS

I am indebted to Dave Rumelhart for several years of provocative conversations on many of these issues; his contributions permeate the ideas formulated here. Sincere thanks to Jerry Fodor and Zenon Wylshyn for most instructive conversations. Comments on earlier drafts from Geoff Hinton, Mark Fantz, and Dan Lloyd were very helpful, as were pointers from Kathleen Akins. Extended comments on the manuscript by Georges Rey were extremely helpful. I am particularly grateful for a number of insights that Rob Cummins and Denise Dellarosa have generously contributed to this paper.

This research has been supported by NSF grant IST-8609599 and by the Department of Computer Science and Institute of Cognitive Science at the University of Colorado at Boulder.

NOTES

¹ In this target article, when *interpretation* is used to refer to a process, the sense intended is that of computer science: the process of taking a linguistic description of a procedure and executing that procedure.

² Consider, for example, the connectionist symposium at the University of Geneva held Sept. 9, 1986. The advertised program featured Feldman, Minsky, Rumelhart, Sejnowski, and Waltz. Of these five researchers, three were major contributors to the symbolic paradigm for many years (Minsky 1975; Rumelhart 1975; 1980; Waltz 1978).

³ This is an issue that divides connectionist approaches. "Local connectionist models" (e.g., Dell 1985; Feldman 1985; McClelland and Rumelhart 1981; Rumelhart and McClelland 1982; Waltz and Pollack 1985) accept (9), and often deviate significantly from (8a). This approach has been championed by the Rochester connectionists (Feldman *et al.* 1985). Like the symbolic paradigm, this school favors simple semantics and more complex operations. The processors in their networks are usually more powerful than those allowed by (8); they are often like digital computers running a few lines of simple code. ("If there is a 1 on this input line then do *X* else do *Y*," where *X* and *Y* are quite different simple procedures; e.g., Shastri 1985.) This style of connectionism, quite different from the subsymbolic style, has much in common with techniques of traditional computer science for "parallelizing" serial algorithms by decomposing them into routines that can run in parallel, often with certain synchronization points built in. The grain size of the Rochester parallelism, although large compared to the subsymbolic paradigm, is small compared to standard parallel programming: The processors are allowed only a few internal states and can transmit only a few different values (Feldman and Ballard 1982).

⁴ As indicated in the introduction, a sizeable sample of research that by and large falls under the subsymbolic paradigm can be found in the books, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*: Rumelhart, McClelland, and the PDP Research Group 1986; McClelland, Rumelhart, and the PDP Research

Group 1986. While this work has since come to be labelled "connectionist," the term "PDP" was deliberately chosen to distinguish it from the localist approach, which had previously adopted the name "connectionist" (Feldman and Ballard 1982).

⁵ The phrase is Roger Schank's, in reference to "parallel processing" (Waldrop 1984). Whether he was referring to connectionist systems I do not know; in any event, I don't mean to imply that the grounds for his comment are addressed here.

⁶ In this section the disclaimer in the introduction is particularly relevant: The arguments I offer are not intended to represent a consensus among connectionists.

⁷ For example, two recently discovered learning rules that allow the training of hidden units, the Boltzmann machine learning procedure (Hinton and Sejnowski 1983a) and the back-propagation procedure (Rumelhart, Hinton and Williams 1986), both involve introducing computational machinery that is motivated purely mathematically; the neural counterparts of which are so far unknown (unit-by-unit connection strength symmetry, alternating Hebbian and antiHebbian learning, simulate annealing, and backwards error propagation along connections of identical strength to forward activation propagation).

⁸ A notable exception is Touretzky and Hinton 1985.

⁹ Furthermore, when a network makes a mistake, it can be told the correct answer, but it cannot be told the precise rule it violated. Thus it must assign blame for its error in an undirected way. It is quite plausible that the large amount of training currently required by subsymbolic systems could be significantly reduced if blame could be focused by citing violated rules.

¹⁰ There is a trade-off between the number of goal conditions one chooses to attribute to a system, and the corresponding range of tolerable environmental conditions. Considering a large variety of environmental conditions for a river, there is only the increase the corresponding goal repertoire. A river can "flow downhill" goal; by appropriately narrowing the class of conditions, one can meet the goal of carrying messages from *A* to *B*, if *A* and *B* are appropriately restricted. But a homing pigeon can meet this goal over a much greater variety of situations.

¹¹ Consider a model that physicists like to apply to "neural nets" – the *spin glass* (Toulouse *et al.* 1986). Spin glasses seem relevant because they are dynamical systems in which the interactions of the variables ("spins") are spatially inhomogeneous. But a spin glass is a system in which the interactions between spins are random variables that all obey the *same* probability distribution *p*: The system has *homogeneous inhomogeneity*. The analysis of spin glasses relates the properties of *p* to the bulk properties of the medium as a whole; the analysis of a single spin subject to a particular set of inhomogeneous interactions is regarded as quite meaningless, and techniques for such analysis are not generally developed.

¹² This problem is closely related to the localization of a failure of veridicality in a scientific theory. Pursuing the remarks of Section 2.1, scientific theories can be viewed as cognitive systems, indeed ones having the prediction goal. Veridicality is a property of a scientific theory as a whole, gauged ultimately by the success or failure of the theory to meet the prediction goal. The veridicality of abstract representations in a theory derives solely from their causal role in the accurate predictions of observable representations.

¹³ For example, in both harmony theory and the Boltzmann machine, discrete units have typically been used because (1) discrete units simplify both analysis and

simulation; (2) for the quadratic harmony or energy functions that are being optimized, it can be proved that no optima are lost by simplifying to binary values; (3) these models' stochastic search has a "jumpy" quality to it anyway. These, at least, are the *computational* reasons for discrete units; in the case of the Boltzmann machine, the discrete nature of action potentials is also cited as a motivation for discrete units (Hinton *et al.* 1984).

¹⁴ Alternatively, if the original harmony/Boltzmann approach is extended to include nonquadratic harmony/energy functions, nonbinary optima appear, so again one switches to continuous units (Derthick, in progress; Smolensky, in progress).

¹⁵ Note that these (procedural) "productions" that occur in intuitive processing are very different from the (declarative) production rules of Section 6 that occur in conscious rule application.

REFERENCES

- Ackley, D.H., Hinton, G.E. and Sejnowski, T.J. (1985) A learning algorithm for Boltzmann machines. *Cognitive Science* 9:147-69.
- Anderson, J.R. (1981) *Cognitive skills and their acquisition*. Erlbaum.
- Anderson, J.R. (1983) *The architecture of cognition*. Harvard University Press.
- Anderson, J.R. (1985) *Cognitive Science* 9(1): Special issue on connectionist models and their applications.
- Ballard, D.H. (1966) Cortical connections and parallel processing: Structure and function. *Behavioral and Brain Sciences* 9:67-120.
- Ballard, D.H. and Hayes, P.J. (1984) Parallel logical inference. *Proceedings of the Sixth Conference of the Cognitive Science Society*.
- Cohen, M.S. (1986) Design of a new medium for volume holographic information processing. *Applied Optics* 14:2288-94.
- Crick, F. and Asanuma, C. (1986) Certain aspects of the anatomy and physiology of the cerebral cortex. In: *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 2: *Psychological and biological models*, ed. J.L. McClelland, D.E. Rumelhart and the PDP Research Group. MIT Press/Bradford Books.
- Dell, G.S. (1985) Positive feedback in hierarchical connectionist models: Applications to language production. *Cognitive Science* 9:3-23.
- Derthick, M.A. (1986) *A connectionist knowledge representation system*. Thesis proposal, Computer Science Department, Carnegie-Mellon University.
- Feldman, J.A. (1961) A connectionist model of visual memory. In: *Parallel models of associative memory*, ed. C.E. Hinton and J.A. Anderson. Erlbaum.
- Feldman, J.A. (1985) Four frames suffice: A provisional model of vision and space. *Behavioral and Brain Sciences* 8:265-89.
- Feldman, J.A. (1986) Neural representation of conceptual knowledge. Technical Report 189, Department of Computer Science, University of Rochester.
- Feldman, J.A. and Ballard, D.H. (1982) Connectionist models and their properties. *Cognitive Science* 6:205-54.
- Feldman, J.A., Ballard, D.H., Brown, C.M. and Dell, G.S. (1985) Rochester connectionist papers: 1979-1985. Technical Report 172, Department of Computer

- Science, University of Rochester.
- Fodor, J.A. (1975) *The language of thought*. Harvard University Press.
- Fodor, J.A. (1987) Why there still has to be language of thought. In: *Psychosemantics*, ed. J.A. Fodor. MIT Press/Bradford Books.
- Fodor, J.A. and Pylyshyn, Z.W. (1988) Connectionism and cognitive architecture: A critical analysis. *Cognition* 28:3-71.
- Geman, S. and Geman, D. (1964) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6:721-41.
- Haugeland, J. (1978) The nature and plausibility of cognitivism. *Behavioral and Brain Sciences* 1:215-26.
- Hebb, D.O. (1949) *The organization of behavior*. Wiley.
- Hinton, G.E. and Anderson, J.A. eds. (1961) *Parallel models of associative memory*. Erlbaum.
- Hinton, G.E., McClelland, J.L. and Rumelhart, D.E. (1986) Distributed representations. In: *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 1: *Foundations*, ed. J.L. McClelland, D.E. Rumelhart and the PDP Research Group. MIT Press/Bradford Books.
- Hinton, G.E. and Sejnowski, T.J. (1983a) Analyzing cooperative computation. *Proceedings of the Fifth Annual Conference of the Cognitive Science Society*.
- Hinton, G.E. and Sejnowski, T.J. (1983b) Optimal perceptual inference. *Proceedings of the I.E.E.E. Conference on Computer Vision and Pattern Recognition*.
- Hinton, G.E. and Sejnowski, T.J. (1986) Learning and relearning in Boltzmann machines. In: *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 1: *Foundations*, ed. J.L. McClelland, D.E. Rumelhart and the PDP Research Group. MIT Press/Bradford Books.
- Hofstadter, D.R. (1979), *Godel, Escher, Bach: An eternal golden braid*. Basic Books.
- Hofstadter, D.R. (1985a) Variations on a theme as the crux of creativity. In: *Metamagical themas*. Basic Books.
- Hofstadter, D.R. (1985b) Waking up from the Boolean dream, or, subcognition as computation. In: *Metamagical themas*. Basic Books.
- Hopcroft, J.E. and Ullman, J.D. (1979) *Introduction to automata theory, languages, and computation*. Addison-Wesley.
- Jordan, M.I. (1986) Attractor dynamics and parallelism in a connectionist sequential machine. *Proceedings of the Eighth Meeting of the Cognitive Science Society*.
- Kant, I. (1787/1963) *Critique of pure reason*. N. Kemp Smith, trans., 2nd ed. Macmillan.
- Larkin, J.H., McDermott, J., Simon, D.P. and Simon, H.A. (1980) Models of competence in solving physics problems. *Cognitive Science* 4:317-45.
- Lashley, K. (1950) In search of the engram. In: *Psychological mechanisms in animal behavior*, Symposia of the Society for Experimental Biology, No. 4, Academic Press.
- Lewis, C.H. (1978) *Production system models of practice effects*.
- McClelland, J.L. and Rumelhart, D.E. (1981) An interactive activation model of context effects in letter perception: Part 1. An account of the basic findings. *Psychological Review* 88:375-407.

- McClelland, J.L., Rumelhart, D.E. and the PDP Research Group (1986) *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 2: *Psychological and biological models*. MIT Press/Bradford Books.
- Minsky, M. (1963) Steps toward artificial intelligence. In: *Computers and thought*, ed. E.A. Feigenbaum and J. Feldman. McGraw-Hill.
- Minsky, M. (1975) A framework for representing knowledge. In: *The psychology of computer vision*, ed. P.H. Winston, McGraw-Hill.
- Minsky, M. and Papert, S. (1969) *Perceptrons* MIT Press.
- Newell, A. (1980) Physical symbol systems. *Cognitive Science* 4:135-83.
- Newell, A. and Simon, H.A. (1972) *Human problem solving*. Prentice Hall.
- Pearl, J. (1985) Bayesian networks: A model of self-activated memory for evidential reasoning. *Proceedings of the Seventh Conference of the Cognitive Science Society*.
- Pylyshyn, Z.W. (1984) *Computation and cognition: Toward a foundation for cognitive science*. MIT Press/Bradford Books.
- Riley, M.S. and Smolensky, P. (1984) A parallel model of (sequential) problem solving. *Proceedings of the Sixth Annual Conference of the Cognitive Science Society*.
- Rumelhart, D.E. (1975) Notes on a schema for stories. In: *Representation and understanding*, ed. D.G. Bobrow and A. Collins. Academic Press.
- Rumelhart, D.E. (1980) Schemata: The building blocks of cognition. In: *Theoretical issues in reading comprehension*, ed. R. Spiro, B. Bruce and W. Brewer. Erlbaum.
- Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) Learning and internal representations by error propagation. In: *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 1: *Foundations*, ed. J.L. McClelland, D.E. Rumelhart and the PDP Research Group. MIT Press/Bradford Books.
- Rumelhart, D.E. and McClelland, J.L. (1982) An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some texts and extensions of the model. *Psychological Review* 89:60-94.
- Rumelhart, D.E. and McClelland, J.L. (1986) On learning the past tense of English verbs. In: *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 2: *Psychological and biological models*, ed. J.L. McClelland, D.E. Rumelhart and the PDP Research Group. MIT Press/ Bradford Books.
- Rumelhart, D.E., McClelland, J.L. and the PDP Research Group (1986) *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 1: *Foundations*. MIT Press/Bradford Books.
- Rumelhart, D.E. Smolensky, P., McClelland, J.L. and Hinton, G.E. (1986) Schemata and sequential thought processes in parallel distributed processing models. In: *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 2: *Psychological and biological models*, ed. J.L. McClelland, D.E. Rumelhart and the PDP Research Group. MIT Press/ Bradford Books.
- Sejnowski, T.J. (1976) On the stochastic dynamics of neuronal interactions. *Biological Cybernetics* 22:203-11.
- Sejnowski, T.J. and Rosenberg, C.R. (1986) NETtalk: A parallel network that learns to read aloud. Technical Report JHU/EECS-86/01, Department of Electrical Engineering and Computer Science, John Hopkins University.

- Shastri, L. (1985) Evidential reasoning in semantic networks: A formal theory and its parallel implementations. Technical Report TR 166, Department of Computer Science, University of Rochester.
- Shepard, R.N. (1962) The analysis of proximities: Multidimensional scaling with an unknown distance function. I and II. *Psychometrika* 27:125–40, 219–46.
- Smolensky, P. (1983) Schema selection and stochastic inference in modular environments. *Proceedings of the National Conference on Artificial Intelligence*.
- Smolensky, P. (1984a) Harmony theory: Thermal parallel models in a computational context. In: *Harmony theory: Problem solving, parallel cognitive models, and thermal physics*, ed. P. Smolensky and M.S. Riley. Technical Report 8404, Institute for Cognitive Science, University of California at San Diego.
- Smolensky, P. (1984b) The mathematical role of self-consistency in parallel computation. *Proceedings of the Sixth Annual Conference of the Cognitive Science Society*.
- Smolensky, P. (1986a) Information processing in dynamical systems: Foundations of harmony theory. In: *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 1: *Foundations*, ed. J.L. McClelland, D.E. Rumelhart and the PDP Research Group. MIT Press/Bradford Books.
- Smolensky, P. (1986b) Neural and conceptual interpretations of parallel distributed processing models. In *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 2: *Psychological and biological models*, ed. J.L. McClelland, D.E. Rumelhart and the PDP Research Group. MIT Press/Bradford Books.
- Smolensky, P. (1986c) Formal modeling of subsymbolic processes: An introduction to harmony theory. In: *Directions in the science of cognition*, ed. N.E. Sharkey. Ellis Horwood.
- Smolensky, P. (1987) On variable binding and the representation of symbolic structures in connectionist systems. Technical Report CU-CS-355-87, Department of Computer Science, University of Colorado at Boulder.
- Smolensky, P. (1988) The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn. *Southern Journal of Philosophy*. Special issue on connectionism and the foundations of cognitive science.
- Toulouse, G., Dehaene, S. and Changeux, J.-P. (1986) A spin glass model of learning by selection. Technical Report, Unite de Neurobiologie Moleculaire, Institut Pasteur, Paris.
- Touretzky, D.S. (1986) BolzCONS: Reconciling connectionism with the recursive nature of stacks and trees. *Proceedings of the Eighth Conference of the Cognitive Science Society*.
- Touretzky, D.S. and Hinton, G.E. (1985) Symbols among the neurons: Details of a connectionist inference architecture. *Proceedings of the International Joint Conference on Artificial Intelligence*.
- Turing, A. (1936) On computable numbers, with an application to Entscheidungs problem. *Proceedings of the London Mathematical Society* (Ser. 2) 42:230–65 and 43:544–46.
- Waldrop, M.M. (1984) Artificial intelligence in parallel. *Science* 225:608–10.
- Waltz, D.L. (1978) An English language question answering system for a large relational database. *Communications of the Association for Computing Machinery*

- 21:526–39.
- Waltz, D.L. and Pollack, J.B. (1985) Massively parallel parsing: A strongly interactive model of natural language interpretation. *Cognitive Science* 9:51–74.

PART III

REPRESENTATIONAL CONCEPTIONS

SEMANTICS, WISCONSIN STYLE

There are, of course, two kinds of philosophers. One kind of philosopher takes it as a working hypothesis that belief/desire psychology (or, anyhow, *some* variety of propositional attitude psychology) is the best theory of the cognitive mind that we can now envision; hence that the appropriate direction for psychological research is the construction of a belief/desire theory that is empirically supported and methodologically sound. The other kind of philosopher takes it that the entire apparatus of propositional attitude psychology is conceptually flawed in irremediable ways; hence that the appropriate direction for psychological research is the construction of alternatives to the framework of belief/desire explanation. This way of collecting philosophers into philosopher-kinds cuts across a number of more traditional, but relatively superficial, typologies. For example, eliminativist behaviorists like Quine and neurophiles like the Churchlands turn up in the same basket as philosophers like Steve Stich, who think that psychological states are computational and functional all right, but not intentional. Dennett is probably in that basket too, along with Putnam and other (how should one put it?) dogmatic relativists. Whereas, among philosophers of the other kind one finds a motley that includes, very much *inter alia*, reductionist behaviorists like Ryle and (from time to time) Skinner, radical individualists like Searle and Fodor, mildly radical anti-individualists like Burge, and, of course, all cognitive psychologists except Gibsonians.

Philosophers of the first kind disagree with philosophers of the second kind about many things besides the main issue. For example, they tend to disagree vehemently about who has the burden of argument. However – an encouraging sign – recent discussion has increasingly focused upon one issue as the crux par excellence on which the resolution of the dispute must turn. The point about propositional attitudes is that they are *representational* states: Whatever else a belief is, it is a kind of thing of which semantic evaluation is appropriate. Indeed, the very individuation of beliefs proceeds via (oblique) reference to the states of affairs that determine their semantic value; the

belief that it is raining is essentially the belief whose truth or falsity depends on whether it is raining. Willy nilly, then, the friends of propositional attitudes include only philosophers who think that serious sense can be made of the notion of representation (de facto, they tend to include *all* and only philosophers who think this). I emphasize that the notion of representation is crucial for every friend of propositional attitudes, not just the ones (like, say, Field, Harman and Fodor) whose views commit them to quantification over symbols in a mental language. Realists about propositional attitudes are *ipso facto* Realists about representational states. They must therefore have *some* view about what it is for a state to *be* representational even if (like, say, Loar and Stalnaker) they are agnostic about, or hostile towards, identifying beliefs and desires with sentences in the language of thought.

Well, what would it be like to have a serious theory of representation? Here, too, there is some consensus to work from. The worry about representation is above all that the semantic (and/or the intentional) will prove permanently recalcitrant to integration in the natural order; for example, that the semantic/intentional properties of things will fail to supervene upon their physical properties. What is required to relieve the worry is therefore, at a minimum, the framing of *naturalistic* conditions for representation. That is, what we want at a minimum is something of the form '*R represents S*' is true iff *C* where the vocabulary in which condition *C* is couched contains neither intentional nor semantical expressions.^{1,2}

I haven't said anything, so far, about what *R* and *S* are supposed to range over. I propose to say as little about this as I can get away with, both because the issues are hard and disputatious and because it doesn't, for the purposes of this paper, matter much how they are resolved. First, then, I propose to leave it open which things *are* representations and how many of the things that qualify a naturalistic theory should cover. I assume only that we must have a naturalistic treatment of the representational properties of the propositional attitudes; if propositional attitudes are relations to mental representations, then we must have a naturalistic treatment of the representational properties of the latter.

In like spirit, I propose to leave open the ontological issues about the possible values of *S*. The paradigmatic representation relation I have in mind holds between things of the sorts that have truth values and things of the sorts by which truth values are determined. I shall usually refer to

the latter as "states of affairs", and I'll use '-ing nominals' as canonical forms for expressing them (eg., 'John's going to the store'; 'Mary's kissing Bill'; 'Sam's being twelve years old next Tuesday'). Since the theories we'll discuss hold that the relations between a representation and what it represents are typically causal, I shall assume further that *S* ranges over kinds of things that can *be* causes.

Last in this list of things that I'm not going to worry about is type¹ token ambiguities. A paradigm of the relation we're trying to provide a theory for is the one that holds between my present, occurrent belief that Reagan is president and the state of affairs consisting of Reagan's being President. I assume that this is a relation between tokens; between an individual belief and an individual state of affairs. But I shall also allow talk of relations between representation *types* and state of affair *types*; the most important such relation is the one that holds when tokens of a situation type cause, or typically cause, tokenings of a representation type. Here again there are ontological deep waters; but I don't propose to stir them up unless I have to.

OK, let's go. There are, so far as I know, only two sorts of naturalistic theories of the representation relation that have ever been proposed. And at least one of these is certainly wrong. The two theories are as follows: that *C* specifies some sort of *resemblance* relation between *R* and *S*; and that *C* specifies some sort of *causal* relation between *R* and *S*.³ The one of this pair that is certainly wrong is the resemblance theory. For one thing, as everybody points out, resemblance is a symmetrical relation and representation isn't; so resemblance can't *be* representation. And, for another, resemblance theories have troubles with the *singularity* of representation. The concept *tiger* represents *all tigers*; but the concept *this tiger* represents only this one. There must be (possible) tigers that resemble this tiger to any extent you like, and if resemblance is sufficient for representation, you'd think the concept *this tiger* should represent those tigers too. But it doesn't, so again resemblance can't be sufficient for representation.

All this is old news. I mention it only to indicate some of the ways in which the idea of a causal theory of representation is *prima facie* attractive, and succeeds where resemblance theories fail. (1) Causal relations are natural relations if *anything* is. You might wonder whether resemblance is part of the natural order (or whether it's only, as it were, in the eye of the beholder). But to wonder that about causation is to wonder whether there *is* a natural order. (2) Causation, unlike resem-

blance, is nonsymmetric, (3) Causation is par excellence, a relation among *particulars*. Tiger *a* can resemble tiger *b* as much as you like, and it can still be tiger *a* and not tiger *b* that caused this set of tiger prints. Indeed, if it was tiger *a* that caused them, it follows that tiger *b* didn't (assuming, of course, that tiger *a* is distinct from tiger *b*).

Well, in light of all this, several philosophers who are sympathetic towards propositional attitudes have recently been playing with the idea of a causal account of representation (see, particularly, Stampe (1975; 1977), Dretske (1981; forthcoming) and Fodor (forthcoming). Much of this has been going on at the University of Wisconsin, hence the title of this essay.) My present purpose is to explore some consequences of this idea. Roughly, here's how the argument will go: causal theories have trouble distinguishing the conditions for *representation* from the conditions for *truth*. This trouble is intrinsic; the conditions that causal theories impose on representation are such that, when they're satisfied, *misrepresentation* cannot, by that very fact, occur. Hence, causal theories about how propositional attitudes represent have Plato's problem to face: how is false belief possible? I'll suggest that the answer turns out to be that, in a certain sense, it's not, and that this conclusion may be more acceptable than at first appears.

I said I would argue for all of that; in fact I'm going to do less. I propose to look at the way the problem of misrepresentation is handled in the causal theories that Stampe and Dretske have advanced; and I really *will* argue that their treatments of misrepresentation don't work. This exercise should make it reasonably clear why misrepresentation is so hard to handle in causal theories generally. I'll then close with some discussion of what we'll have to swallow if we choose to bite the bullet. The point of all this, I emphasize, is *not* to argue against causal accounts of representation. I think, in fact, that something along the causal line is the best hope we have for saving intentionalist theorizing, both in psychology and in semantics. But I think too that causal theories have some pretty kinky consequences, and it's these that I want to make explicit.

To start with, there are, strictly speaking, *two* Wisconsin theories about representation; one that's causal and one that's epistemic. I propose to give the second pretty short shrift, but we'd better have a paragraph or two.

The basic idea of (what I shall call) an epistemic access theory is that *R* represents *S* if you can find out about *S* from *R*.⁴ So, for example,

Dretske says (EB, 10) “A message . . . carries information about X to the extent to which one could learn (come to know) something about X from the message.” And Stampe says (S&T 223): “An object will represent or misrepresent the situation . . . only if it is such as to enable one to come to know the situation, i.e., what the situation is, should it be a faithful representation.”

Now, generally speaking, if representation requires that S cause R , then it will of course be possible to learn about R by learning about S ; inferring from their effects is a standard way of coming to know about causes. So, depending on the details, it’s likely that an epistemic account of representation will be satisfied whenever a causal one is. But there is no reason to suppose that the reverse inference holds, and we’re about to see that epistemic accounts have problems to which the causal ones are immune.

(1) The epistemic access story (like the resemblance story) has trouble with the nonsymmetry of representation. You can find out about the weather from the barometer, but you can also find out about the barometer from the weather since, if it’s storming, the barometer is likely to be low. Surely the weather doesn’t represent the barometer, so epistemic access can’t be sufficient for representation.

(2) The epistemic story (again like the one about resemblance) has trouble with the singularity of representation. What shows this is a kind of case that Stampe discusses extensively in TCTL. Imagine a portrait of, say, Chairman Mao. If the portrait is faithful, then we can infer from properties of the picture to properties of the Chairman (e.g., if the portrait is faithful, then if it shows Mao as bald, then we can learn *from the portrait that* Mao is bald). The trouble is, however, that if Mao has a Doppelgänger and we know he does, then we can *also* learn from the portrait that Mao’s Doppelgänger is bald. But the portrait is of Mao and not of his Doppelgänger for all that.

Dretske has a restriction on his version of the epistemic access theory that is, I expect, intended to cope with the singularity problem; he allows that a message carries information about X only if a “suitably equipped but *otherwise ignorant* receiver” could learn about X from the message (EB 10, my emphasis). I imagine the idea is that, though we could learn about Mao’s Doppelgänger from Mao’s portrait, we couldn’t do so *just from the portrait alone*; we’d also have to use our knowledge that Mao has a Doppelgänger. I doubt, however, that this further condition can really be enforced. What Dretske has to face is, in effect,

the Dreaded Collateral Information Problem; i.e., the problem of how to decide when the knowledge that we use to interpret a symbol counts as knowledge about the symbol, and when it counts as collateral knowledge. This problem may seem self-solving in the case of *pictures* since we have a pretty good pretheoretical notion of which properties of a picture count as the pictorial ones. But in the case of, e.g., linguistic symbols, it's very far from evident how, or even whether, the corresponding distinction can be drawn. If I say to you "John is thirty two", you can learn something reliable about John's age from what I said. But, of course, you can also learn something reliable about John's *weight* (e.g., that he weighs more than a gram). It may be possible to discipline the intuition that what you learn about John's age you learn just from the symbol and what you learn about his weight you learn from the symbol plus background information. But drawing that distinction is notoriously hard and, if the construal of representation depends on our doing so, we are in serious trouble.

(3) Epistemic theories have their own sorts of problems about misrepresentation. Stampe says,

An object will represent or misrepresent the situation . . . only if it is such as to enable one to come to know the situation, i.e., what the situation is, should it be a faithful representation. If it is not faithful, it will misrepresent the situation. That is, one *may* not be able to tell from it what the situation is, despite the fact that it is a representation of the situation. In either case, it represents the same thing, just as a faithful and an unrecognizable portrait may both portray the same person.

But, to begin with, the example is perhaps a little question-begging, since it's not clear that the bad portrait represents its sitter *in virtue of* the fact that if it were accurate it would be possible to learn from it how the sitter looks. How, one wonders, could this bare counterfactual determine representation? Isn't it, rather, the other way around; i.e., not that it's a portrait of Mao because (if it's faithful) you can find out about Mao from it, but rather that you can find out about Mao from it (if it's faithful) because it's Mao that it's a portrait of.

To put the same point slightly differently: we'll see that causal theories have trouble saying how a symbol could be tokened and still be false. The corresponding problem with epistemic access theories is that they make it hard to see how a symbol could be *intelligible* and false. Stampe says: "An object will represent or misrepresent the situation . . . only if it is such as to enable one to come to know the situation, i.e., what the situation is, should it be a faithful represen-

tation.” (S&T 223). Now, there is a nasty scope ambiguity in this; viz., between:

- (a) if *R* is faithful (you can tell what the case is); vs.
- (b) you can tell (what the case is if *R* is faithful).

It's clear that it is (a) that Stampe intends; ((b) leads in the direction of a possible world semantics, which is where Stampe explicitly doesn't want to go; see especially SAT, circa p. 224). So, consider the symbol “Tom is Armenian”, and let's suppose the fact – viz., the fact in virtue of which that symbol has its truth value – is that Tom is Swiss. Then Stampe wants it to be that what the symbol represents (i.e., *misrepresents*) is Tom's being Swiss; *that's* the fact to which, if it were faithful, the symbol would provide epistemic access.

Now, to begin with, this counterfactual seems a little queer. What, precisely, would it be *like* for “Tom is Armenian” to be faithful to the fact it (mis)represents – viz., to the fact that Tom is Swiss? Roughly speaking, you can make a false sentence faithful either by changing the world or by changing the sentence; but neither will do the job that Stampe apparently wants done.

(1) Change the world: make it be that Tom is Armenian. The sentence is now faithful, but to the wrong fact. That is, the fact that it's now faithful to isn't the one that it (mis)represented back when it used to be untrue; that, remember, was the fact that Tom is Swiss.

(2) Change the sentence: make it *mean* that Tom is Swiss. The sentence is now faithful to the fact that it used to (mis)represent. But is the counterfactual intelligible? Can we make sense of talk about what a sentence would represent if it – the very same sentence – meant something different? And, if meaning can change while what is represented stays the same, in what sense does a theory of representation constitute a theory of meaning?

Problems, problems. Anyhow, the main upshot is clear enough, and it's one that Stampe accepts. According to the epistemic access story, when a symbol *misrepresents*, “one *may* not be able to tell from it what the situation is, despite the fact that it is a representation of the situation”. Here not being “able to tell what the situation is” doesn't mean not being able to tell what it is that's *true* in the situation; it means not being able to tell *what situation it is that the symbol represents*. You can't tell, for example, that the symbol “Tom is Armenian” represents

Tom's being French unless you happen to know Tom's nationality.

It may be supposed that Stampe could disapprove of this along the following lines: you *can*, in one sense, tell what "Tom is Armenian" represents even if you don't know that Tom is Swiss. For, you can know that "Tom is Armenian" represents Tom's nationality (i.e., that if it's faithful it provides epistemic access to his nationality) even if you don't know what Tom's nationality is. I think this is OK, but you buy it at a price: On this account, knowing what a symbol *represents* (what it provides epistemic access to) can't be equated with knowing what the symbol *means*. Notice that though "Tom is Armenian" has the property that if it's faithful it provides epistemic access to Tom's nationality, so too do a scillion other, nonsynonymous sentences like "Tom is Dutch", "Tom is Norwegian", "Tom is Swiss", and so forth. To put the same point another way, on the present construal of Stampe's account, what a truth valuable symbol represents isn't, in general, its truth condition. (The truth condition of a symbol is the state of affairs which, if it obtains, would make the symbol true; and what would make "Tom is Armenian" true is Tom's being Armenian, not Tom's being Swiss.) Correspondingly, what you can know about "Tom is Armenian" if you don't know that Tom is Swiss is not what its truth condition is, but only what it represents; viz., that it represents Tom's nationality. This means that Stampe has either to give up on the idea that understanding a symbol is knowing what would make it true, or develop a reconstruction of the notion of truth condition as well as a reconstruction of the notion of representation. Neither of these alternatives seems particularly happy.

There's more to be said about the epistemic approach to representation; but let's, for present purposes, put it to one side. From here on, only causal accounts will be at issue.

The basic problem for causal accounts is easy enough to see. Suppose that *S* is the truth condition of *R* in virtue of its being the cause of *R*. Now, causation is different from resemblance in the following way; a symbol can (I suppose) resemble something merely possible; it's OK for a picture to be a picture of a unicorn. But, surely, no symbol can be an effect of something merely possible. If *S* causes *R*, then *S* obtains. But if *S* obtains and *S* is the truth condition of *R*, it looks as though *R* has to be true; being true just *is* having truth conditions that obtain. So it looks like this: a theory that numbers *causation* among the relations in virtue of which a representation has its truth conditions is going to allow

truth conditions to be assigned only when they're satisfied. I don't say that this argument is decisive; but I do say – and will now proceed to argue – that Wisconsin semantics hasn't thus far found a way around it.

I'll start with Dretske's treatment of the misrepresentation problem in *Knowledge And the Flow of Information*. The crucial passage is on pp. 194–195. Here is what Dretske says:

In the learning situation special care is taken to see that incoming signals have an intensity, a strength, sufficient unto delivering the required piece of information *to* the learning subject . . . Such precautions are taken in the learning situation . . . in order to ensure that an internal structure is developed with the information that *s* is *F* . . . But once we have meaning, once the subject has articulated a structure that is selectively sensitive to information about the *F*-ness of things, instances of this structure, tokens of this type, can be triggered by signals that *lack* the appropriate piece of information . . . We (thus) have a case of misrepresentation – a token of a structure with a false content. We have, in a word, meaning with truth. (Emphasis Dretske's.)

All you need to remember to understand this well enough for present purposes is (1) that Dretske's notion of information is fundamentally that of counterfactual supporting correlation (i.e., that objects of type *R* carry information about states of affairs of type *S* to the extent that tokenings of the type *S* are nomically responsible for tokenings of the type *R*). And (2) that the tokening of a representation carries the information that *s* is *F* in *digital* form if and only if the information that *s* is *F* is the most specific information that that tokening carries about *s*. Roughly speaking, the pretheoretic notion of the *content* of a representation is reconstructed as the information that the representation digitalizes.

Now then: how does *misrepresentation* get into the picture? There is, of course, no such thing as *misinformation* on Dretske's sort of story. Information is correlation and though correlations can be better or worse – more or less reliable – there is no sense to the notion of a *miscorrelation*: hence there is nothing, so far, to build the notion of misrepresentation out of.

The obvious suggestion would be this: suppose *Rs* are nomically correlated with – hence carry information about – *Ss*; then, as we've seen, given the satisfaction of further (digitalization) conditions, we can treat *Rs* as representations of *Ss*: *S* is the state of affairs type that symbols of the *R* type represent. But suppose that, from time to time, tokenings of *R* are brought about (not by tokenings of *S* but) in some *other* way. Then these, as one might say, 'wild' tokenings would count

as *misrepresentations*: for, on the one hand, they have the content that *S*; but, on the other hand, since it isn't the fact that *S* that brings about their tokening the content that they have is false. *Some* sort of identification of misrepresentations with etiologically wild tokenings is at the heart of all causal accounts of misrepresentation.

However, the crude treatment just sketched clearly won't do: it is open to an objection that can be put like this: If there are wild tokenings of *R*, it follows that the nomic dependence of *R* upon *S* is imperfect; some *R*-tokens – the wild ones – are *not* caused by *S* tokens. Well, but clearly they are caused by *something*; i.e., by something that is, like *S*, sufficient but not necessary for bringing *R*s about. Call this second sort of sufficient condition the tokening of situations of type *T*. Here's the problem: *R* represents the state of affairs with which its tokens are causally correlated. Some representations of type *R* are causally correlated with states of affairs of type *S*; some representations of type *R* are causally correlated with states of affairs of type *T*. So it looks as though what *R* represents is not either *S* or *T*, but rather the disjunction (*S* \vee *T*): The correlation of *R* with the disjunction is, after all, *better than* its correlation with either of the disjuncts and, ex hypothesi, correlation makes information and information makes representation. If, however, what *R*s represent is not *S* but (*S* \vee *T*), then tokenings of *R* that are caused by *T* aren't, *after all*, *wild tokenings* and our account of misrepresentation has gone West.

It is noteworthy that this sort of argument – which, in one form or other, will be with us throughout the remainder of this essay – seems to be one that Dretske himself accepts. The key assumption is that, *ceteris paribus*, if the correlation of a symbol with a disjunction is better than its correlation with either disjunct, it is the disjunction, rather than either disjunct, that the symbol represents. This is a sort of “principle of charity” built into causal theories of representation: ‘so construe the content of a symbol that what it is taken to represent is what it correlates with *best*’. Dretske apparently subscribes to this. For example, in EB (circa p. 17) he argues that, for someone on whose planet there is both XYZ and H₂O but who learns the concept *water* solely from samples of the former, the belief that such and such is water is the belief that it is that it is *either* H₂O *or* XYZ. This seems to be charity in a rather strong form: *R* represents a disjunction even if all tokenings of *R* are caused by the satisfaction of the *same* disjunct, so long as satisfaction of the other disjunct *would have caused R tokenings had*

they happened to occur. I stress this by way of showing how much the counterfactuals count; Dretske's conditions on representation are intensional (with an 's'); they constrain the effects of counterfactual causes.

To return to Dretske's treatment of misrepresentation: his way out of the problem about disjunction is to enforce a strict distinction between what happens in the learning period and what happens after. Roughly, the correlations that the learning period establish determine what *R* represents; and the function of the Teacher is precisely to insure that the correlation so established is a correlation of *R* tokens with *S* tokens. It may be that *after* the learning period, *R* tokens are brought about by something *other than S* tokens; if so, these are wild tokenings of *R* and their contents are false.

This move is ingenious but hopeless. Just for starters, the distinction between what happens in the learning period and what happens thereafter surely isn't principled; there is no time after which one's use of a symbol stops being merely shaped and starts to be, as it were, in earnest. Perhaps idealization will bear some of this burden, but it's hard to believe that it could yield a notion of learning period sufficiently rigorous to underwrite the distinction between truth and falsity; which is, after all, precisely what's at issue. Second, if Dretske does insist upon the learning period gambit, he limits the applicability of his notion of misrepresentation to *learned* symbols. This is bad for me because it leaves us with no way in which innate information could be false; and it's bad for him because it implies a basic dichotomy between *natural* representation (smoke and fire; rings in the tree and the age of the tree) and the intentionality of mental states.

All of that, however, is mere limbering up. The real problem about Dretske's gambit is internal; it just doesn't work. Consider a trainee who comes to produce *R* tokens in *S* circumstances during the training period. Suppose, for simplification, that the correlation thus engendered is certainly nomic, and that *S* tokenings are elicited by *all and only R* tokenings during training: error-free learning. Well, time passes, a whistle blows (or whatever), and the training period comes to an end. At some time later still, the erstwhile trainee encounters a tokening of a *T* situation (*T* not equal to *S*) and produces an *R* in causal consequence. The idea is, of course, that this *T*-elicited tokening of *R* is ipso facto wild and, since it happens after the training period ended, it has the (false) content that *S*.

But, as I say, this won't work: it ignores relevant counterfactuals. Imagine, in particular, what *would have* happened if a token of situation type *T* *had* occurred during the training period. Presumably what would have happened is that it would have elicited a tokening of *R*. After all, tokenings of *T* are assumed to be sufficient to cause *R* tokenings *after* training; that's the very assumption upon which Dretske's treatment of wild *R*-tokenings rests. So we can assume – indeed, we can stipulate – that *T* is a situation which, if it had occurred *during* training, would have been sufficient for *R*. But that means, of course, that if you include the counterfactuals, the correlation that training established is (not between *R* and *S* but) between *R* and the disjunction (*S* \vee *T*). So now we have the old problem back again. If training established a correlation with (*S* \vee *T*) then the content of a tokening of *R* is *that* (*S* \vee *T*). So a tokening of *R* caused by *T* isn't a wild tokening after all; and since it isn't wild it also isn't false. A token with the content (*S* \vee *T*) is, of course, *true* when it's the case that *T*.

There is a sort of way out for Dretske. He could say this: 'The trouble is, you still haven't taken care of *all* the relevant counterfactuals; in particular, you've ignored the fact that if a *T*-tokening has occurred during training and elicited an *R*-tokening *the Teacher would have corrected the R response*. This distinguishes the counterfactual consequences of *T*-elicited *R*-tokens occurring during training from those of *S*-elicited *R*-tokens occurring during training since the latter would not, of course, have been corrected. In the long run, then, it is *these* counterfactuals – ones about what the teacher *would have corrected* – that are crucial; *R*s represent *S*s (and not *T*s) because the Teacher would have disapproved of *T*-elicited *R*-responses if they had occurred.'

But I don't think Dretske would settle for this, and nor will I. It's no good for Dretske because it radically alters the fundamental principle of his theory, which is that the character of symbol-to-situation correlations determines the content of a symbol. On this revised view, the essential determinant is not the actual, or even the counterfactual, correlations that hold between the symbol and the world; rather it's the Teacher's pedagogical intentions; specifically, the Teacher's intention to reward only such *R* tokenings as are brought about by *S*s. And it's no good for me because it fails a prime condition upon *naturalistic* treatment of representations; viz, that appeals to intentional (with a 't') states must not figure essentially therein. I shall therefore put this

suggestion of Dretske's to one side and see what else may be on offer.

Let's regroup. The basic problem is that we want there to be conditions for the *truth* of a symbol over and above the conditions whose satisfaction determines what the symbol represents. Now, according to causal theories, the latter – representation determining – conditions include whatever is necessary and sufficient to bring about tokenings of the symbol (including nomically possible counterfactual tokenings.) So the problem is, to put it crudely, if we've already used up all that to establish representation, what more could be required to establish truth?

An idea that circulates in all the texts I've been discussing (including my own) goes like this. Instead of thinking of the representation making conditions as whatever is necessary and sufficient for causing tokenings of the symbol, think of them as whatever is necessary and sufficient for causing such tokenings *in normal circumstances*. We can then think of the wild tokens as being (or, anyhow, as including) the ones which come about when the 'normal conditions' clause is *not satisfied*. This doesn't, of course, get us out of the woods. At a minimum, we still need to show (what is by no means obvious) that for a theory of representations to appeal to normalcy conditions (over and above causal ones) isn't merely question-begging; for example, that you can characterize what it is for the conditions of a tokening to be normal without invoking intentional and/or semantic notions. Moreover, we'll also have to show that appealing to normalcy conditions is a way of solving the disjunction problem; and that, alas, isn't clear either. We commence with the first of these worries.

It is, I think, no accident that there is a tendency in all the texts I've been discussing (again including mine) to introduce normalcy conditions by appeal to examples where *teleology* is in play. For example, to use a case that Dretske works hard, a voltmeter is a device which, under normal conditions, produces an output which covaries (nominally) with the voltage across its input terminals. 'Normal conditions' include that all sorts of constraints on the internal and external environment of the device should be satisfied (e.g., the terminals must not be corroded) but it seems intuitively clear that what the device registers is the voltage and not the voltage together with the satisfaction of the normalcy conditions. If the device reads zero, that means that there's no current flowing, not that *either* there is no voltage flowing *or* the terminals are

corroded.

However, we know this because we know what the device is *for* and we can know what the device is for only because there *is* something that the device is for. The tendency of causal theorists to appeal to teleology for their best cases of the distinction between representation-making causal conditions and mere normalcy conditions is thus unnerving. After all, in the case of artifacts at least, being 'for' something is surely a matter of being *intended* for something. And we had rather hoped to detach the representational from the intentional since, if we can't, our theory of representation ipso facto fails to be naturalistic and the point of the undertaking becomes, to put it mildly, obscure.

There are, it seems, two possibilities. One can either argue that there can be normalcy without teleology (i.e., that there are cases *other than* teleological ones where a distinction between causal conditions and normal conditions can be convincingly drawn); or one can argue that there can be teleology without intentionality (*natural* teleology, as it were) and that the crucial cases of representation rest exclusively upon teleology of this latter kind. Unlike Dretske and Stampe, I am inclined towards the second strategy. It seems to me that our intuitions about the distinction between causal and normal conditions are secure only in the cases where the corresponding intuitions about teleology are secure, and that wherever we *don't* have intuitions about teleology, the disjunction argument seems persuasive.⁵ Let's look at a couple of cases.

Thermometers are OK; given normalcy conditions (e.g., a vacuum in the tube) the nomic covariance between the length of the column and the temperature of the ambient air determines what the device represents. Violate the normalcy conditions and, intuition reports, you get wild readings; i.e., *misrepresentations* of the temperature. But, of course, thermometers are *for* measuring something, and precisely what they're for measuring (viz., the temperature of the ambient air) is what the present analysis treats as a causal (rather than a normalcy) condition. Compare, by way of contrast, the diameter of the coin in my pocket. Fix my body temperature and it covaries with the temperature of the ambient air; fix the temperature of the ambient air, and it covaries with the temperature of my body. I see *no* grounds for saying that one of these things is what it really represents and the other is a normalcy condition (e.g., that the diameters that are affected by body temperature are *misrepresentations* of the air temperature).⁶ In short, where there is no question of teleology it looks as though one's

intuitions about which are the normalcy conditions are unstable. Such examples should make one dubious about the chances for a notion of normalcy that applies in *nonteleological* cases.

Or, consider an example of Stampe's: (CTCLR, 49)

The number of rings in (a tree stump) represents the age of the tree... The causal conditions, determining the production of this representation, are most saliently the climatic conditions that prevailed during the growth of the tree. If these are normal... then one ring will be added each year. Now what is that reading... It is not, for one thing, infallible. There may have been drought years... It is a *conditional* hypothesis: that *if* certain conditions hold, then something's having such and such properties would cause the representation to have such and such properties... Even under those normal conditions, there may be other things that would produce the rings – an army of some kind of borer, maybe, or an omnipotent evil tree demon.

Stampe's analysis of this case rests on his decision to treat the seasonal climatic variations as the causal component of the conditions on representation and the absence of (e.g.) drought, tree borers, evil demons and the rest as normalcy conditions. And, of course, given that decision, it's going to follow from the theory that the tree's rings represent the tree's age and that tree-borer-caused tree ring tokenings are wild (i.e., that they *misrepresent* the tree's age). The worrying question is what, if anything, motivates this decision.

We should do this in several steps. Let's consider a particular case of tree-borer-caused tree ring tokenings. Suppose, for the moment, we agree that the general truth is that a tree's rings represent the tree's age. And suppose we agree that it follows from this general truth that all tree ring tokenings represent the age of the tree that they're tokened in. Well, even given all that it's not obvious what these tree-borer-caused tokenings represent since it's not obvious that they are, in the relevant sense, tree rings.

Perhaps the right way to describe the situation is to say that these things merely *look like* tree rings. Compare the token of "Look upon my works, oh ye mighty, and despair" that the wind traces in the desert sands. This *looks like* a token of an English sentence type (and, of course, if it *were* a token of that sentence type it would be unfaithful, what with there not being anything to look at and all). But it's not a token of that English sentence since it's not a token of *any* sentence. A fortiori, it's not a wild or unfaithful token. Similarly, mutatis mutandis (maybe) with the putative tree rings; they're not wild (unfaithful) representations of the tree's age because, even if all tree rings are

representation of a tree's age, *these aren't tree rings*.

I hope I will be seen not to be merely quibbling. Stampe wants it to come out that tree-borer caused tree rings are wild; that they're misrepresentations of the tree's age. He needs this a lot since this sort of case is Stampe's paradigm example of a distinction between causal conditions and normalcy conditions which doesn't rest on teleology. But I claim that the case doesn't work *even assuming what's yet to be shown*, viz., *that tree rings represent tree age rather than tree-age-plus-satisfaction-of-normalcy-conditions*. For Stampe is assuming a nonquestion begging – hence naturalistic – criterion for something being a token of a representation type. And there isn't one. (Of course, we do have a criterion which excludes the wind token's being a sentence inscription; but that criterion is *nonnaturalistic*, hence unavailable to a causal theorist; it invokes the intentions of the agent who produced the token.)

Now let's look at it the other way. Suppose that these tree-borer caused rings *are* tree rings (by stipulation) and let's ask what they represent. The point here is that even if "under normal conditions, tree rings represent the tree's age" is true, it *still* doesn't follow that *these abnormally formed tree rings represent the tree's age*. Specifically, it doesn't follow that these rings represent the tree's age rather than the tree borer's depredations. (Look closely and you'll see the marks their little teeth left. Do those represent the tree's age too?) This is just the disjunction problem over again, though it shows an interesting wrinkle that you get when you complicate things by adding in normalcy conditions. "If circumstances are normal, *xs are F*" doesn't, of course, tell you about the *F*ness of *xs* when circumstances are *abnormal*. The most you get is a counterfactual, viz., "if circumstances *had been* normal, this *x* would have been *F*." Well, in the present case, if etiological circumstances had been normal, these rings would have represented the tree's age (viz., *accurately*). It doesn't follow that, given the way the etiological circumstances actually were, these rings still represent the tree's age (viz., *inaccurately*). What you need is some reason to suppose that etiologically abnormal (hence wild) rings represent the same thing that etiologically normal rings do. This is precisely equivalent to saying that what you need is a solution to the disjunction problem, and that is precisely what I've been arguing all along that we haven't got.

We *would* have it, at least arguably, if this were a teleological case.

Suppose that there is some mechanism which (not only produces tree rings but) produces tree rings with an end in view. (Tree rings are, let's suppose, Mother Nature's calendar). Then there is a trichotomous distinction between (a) tree rings produced under normal circumstances; (b) wild tree rings (inscribed, for example, when Mother Nature is a little tipsy); and (c) things that look like tree rings but aren't (tree borer's deprivations). This *does* enforce a distinction between representation, misrepresentation and nonrepresentation; not so much because it relativizes representation to *normalcy*, however, but because it relativizes representation to *end-in-view*. The reason that wild tree rings represent the same things as normal ones is that *the wild ones and the normal ones are supposed to serve the same function*. Notice that it's the intensionality of "supposed to" that's doing all the work.

I'm afraid what all this comes to is that the distinction between normal and wild tokens rests – so far at least – on a pretty strong notion of teleology. It's only in the teleological cases that we have any way of justifying the claim that wild tokens represent the same thing that etiologically normal ones do; and it is, as we've seen, that claim on which the present story about misrepresentation rests. How bad is this? Well, for one thing, it's not as bad as if the distinction had turned out to rest on an *intentional* notion. There are, as I remarked above, plausible cases of nonintentional, natural teleology and a naturalistic theory of representation can legitimately appeal to these. On the other hand, if the line of argument we have been exploring is right, then the hope for a *general* theory of representation (one that includes tree rings, for example) is going to have to be abandoned. Tree rings will have to represent only at a remove, via the interests of an observer, since only what has natural teleology can represent absolutely. This is, as a matter of fact, OK with me. For I hold that only sentences in the language of thought represent in, as it were, the first instance; and they represent in virtue of the natural teleology of the cognitive mechanisms. Propositional attitudes represent *qua* relations to sentences in the language of thought. All other representation depends upon the propositional attitudes of symbol users.

Even allowing all this, however, it is arguable that we haven't yet got a notion of misrepresentation robust enough to live with. For we still have this connection between the etiology of representations and their truth values: representations generated in teleologically normal circumstances must be true. Specifically, suppose *M* is a mechanism the

function of which is to generate tokens of representation type *R* in, and only in, tokens of situation type *S*; *M* mediates the causal relation between *Ss* and *Rs*. Then we can say that *M*-produced tokens of *R* are wild when *M* is functioning abnormally; but when *M* is functioning normally (i.e., when its tokening of *R* is causally contingent, in the right way, upon the tokening of *S*) then not only do the tokens of *R* have the content *that S*, but also the contents of these tokens are satisfied, and what the tokens say is true.

Well, consider the application to belief fixation. It looks as though (1) only beliefs with abnormal etiologies can be false, and (2) 'abnormal etiology' will have to be defined with respect to the teleology of the belief-fixing (i.e., cognitive) mechanisms. As far as I can see, this is tantamount to: "beliefs acquired under epistemically optimal circumstances must be true" since, surely, the function of the cognitive mechanisms will itself have to be characterized by reference to the beliefs it *would* cause one to acquire *in* such optimal circumstances. (I take it for granted that we can't, for example, characterize the function of the cognitive mechanisms as the fixation of *true* beliefs because truth is a semantical notion. If our theory of representation is to rest upon the teleology of the cognitive mechanisms, cognitive teleology must itself be describable naturalistically; *viz.*, without recourse to semantic concepts. For an extended discussion of this sort of stuff, see my op cit.)

It appears that we have come all this way only in order to rediscover verificationism. For, I take it, verificationism just *is* the doctrine that truth is what we would believe in cognitively optimal circumstances. Is this simply too shameful for words? Can we bear it? I have three very brief remarks to make. They are, you will be pleased to hear, concluding remarks.

First, *all* Naturalistic theories in semantics, assuming that they are reductive rather than eliminative, have got to hold that there are circumstances, specifiable without resort to semantical notions like truth, reference, correspondence or the like, such that, if a belief is formed *in* those circumstances, then it must be true. Verificationism adds to this only the idea that the circumstances are epistemic (they involve, for example, such idealizations as unrestricted access to the evidence) and that wouldn't seem to be the part that hurts. I guess what I'm saying is: if you're going to be a naturalist, there's no obvious reason not to be a verificationist. (And if you're *not* going to be a naturalist, why are you working on a causal theory of representation?)

The second point is this: verificationism isn't an ontological doctrine. It has usually, in the history of philosophy, been held with some sort of Idealistic malice aforethought, but that surely is an accident and one we can abstract from. The present sort of verificationism defines truth conditions by reference to the function of the cognitive mechanisms. Plausibly, the function of the cognitive mechanisms is to achieve, for the organism, epistemic access to the world. There is no reason on God's green earth why you shouldn't, in parsing that formula, construe "the world" Realistically.

Finally, verificationism isn't incompatible with a correspondance theory of truth. The teleology of the nervous system determines what must be the case if *R* represents *S*; and it follows from the analysis that if *R* represents *S* and the situation is teleologically normal, *S* must be true. This is because what *R* represents is its truth condition, and its truth condition is whatever causes its tokening in teleologically normal situations. But this is entirely compatible with holding that what *makes R* true in teleologically normal situations is that its truth condition obtains; that *R* corresponds, that is to say, to the way that the world is.

I see no way out of this: a causal theory must so characterize representation and normalcy that there is no misrepresentation in normal circumstances. My view is: if that is the price of a workable theory of representation, we ought simply to pay it.

NOTES

¹ Since we haven't any general and satisfactory way of saying which expressions *are* semantical (/intentional), it's left to intuition to determine when a formulation of *C* meets this condition. This will not, however, pose problems for the cases we will examine.

² I said that the formulation of naturalistic conditions for representation is *the least* that the vindication of an intentionalist psychology requires. What worries some philosophers is that there may be no *unique* answer to the question what something represents; e.g., that the representational content of a symbol (belief, etc.) may be *indeterminate* given the totality of physical fact. Notice that settling the question about naturalism doesn't automatically settle this question about determinacy. Even if it proves possible to give naturalistic necessary and sufficient conditions for representation, there might be more than one way to satisfy such conditions, hence more than one thing that *R* could be taken to represent. For purposes of the present paper, however, I propose to put questions about determinacy of representation entirely to one side and focus just on the prospects for naturalism.

³ An example of the former: Propositional attitudes are relations to mental representations; mental representations are Ideas; Ideas are images; and Images represent what

they resemble. I take it that Hume held a view not entirely unlike this.

⁴ In fact, Dretske gives the epistemic analysis as a condition upon '*R carries information about S*' rather than '*R represents S*'. This difference may *make* a difference and I'd have to attend to it if exposition were the goal. In much of what follows, however, I shall be less than sensitive to details of Dretske and Stampe's proposals. What I have in mind to exhibit are certain very pervasive characteristics of causal accounts; ones which I don't *think* can be avoided by tinkering.

⁵ I should add that, though Stampe clearly thinks that you can, in principle, get representation without teleology, cases which turn on functional analysis loom large among his examples: "... one doubts whether statistical normality will get us far in dealing with living systems and with language or generally with matters of teleological nature. Here, I think we shall want to identify fidelity conditions with certain conditions of well functioning, of a functional system." (TCTLR, p. 51).

⁶ Alternatively, you could go the disjunction route and say that the diameter of the coin represents some function of body temperature and air temperature. But this has the familiar consequence of rendering the covariance between *R* and *S* perfect and thus depriving us of examples of wild tokenings.

Dept. of Philosophy,
M.I.T.
Cambridge, MA 02139
U.S.A.

COGNITIVE SCIENCE AND THE PROBLEM OF SEMANTIC CONTENT

1. INTRODUCTION

By 'cognitive science' I mean the branch of cognitive psychology that incorporates the computer model, and that is sometimes known as 'computational psychology' and 'information-processing psychology' as well. My remarks are directed particularly against the version of cognitive science defended in recent writings by Jerry Fodor. By 'the problem of semantic content' I mean roughly the problem of explicating those features of a brain state or process by virtue of which it may properly be said to possess meaning or reference or truth value. Since I am primarily concerned with the account of semantic content implicated in Fodor's version of cognitive science, I shall summarize that account before attempting to formulate the problem in a more precise manner.

The argument of the early portions of the paper is intended to show that Fodor's approach is incapable of providing an adequate answer to the problem of semantic content. I then suggest that this incapacity is symptomatic of a basic confusion between semantic information and information in the technical sense of communication theory. In the final part of the paper I outline an account of semantic content based on the technical sense of information which shows promise of succeeding where Fodor fails.

2. THE BASIC IDEA OF MODERN COGNITIVE THEORY

In Fodor's conception, cognitive science (CS) is closely aligned with artificial intelligence (AI) in that both rely upon the computer model and both stress a "top-down" mode of analysis. This latter is inspired by a strategy of computer programming, which typically begins with a general description of the task to be accomplished, and proceeds by analysis into more and more specific subtasks. Since both AI and CS are concerned with tasks that require intelligence for their successful human performance, and since intelligent human performance invites

description in terms of beliefs and purposes, both approaches begin with descriptions in the intentional idiom.¹ In AI, the goal of subsequent analysis is to break up the task under its intentional description into subtasks within the competence of a digital computer, without concern for how these or alternative subtasks might be performed by the human organism. To be psychologically realistic, on the other hand, an analysis of this sort would have to lead to subtasks that closely match those performed by the brain in the course of actual intelligent activity, and that could be performed in the manner of a digital computer. CS thus imposes the additional requirement that the subtasks be computable by the human nervous system.

Unlike behaviorism, with its exclusion from scientific explanation of reference to processes within the brain of the behaving organism, CS stresses explanation by internal occurrences, and finds its point of departure in our common talk about cognitive attitudes. In particular, the starting point of CS includes generalizations from “folk psychology” (Fodor’s term) in such formats as “seeing that... is a cause of believing that...”, or “believing that... is a cause of saying that...”, where the blanks are to be filled with propositional expressions.² Since explicit reference to beliefs and purposes is unavoidably involved in the very formulation of such generalizations, it follows that CS cannot be conducted in the technical language of physics. Although most advocates of CS consider themselves committed to physicalism as an overall metaphysics, they hold that the “kind terms” (like “belief that so-and-so”) implicated in psychological explanation cannot be expected to coincide with any “kind terms” found in physical science. That is, CS is committed to a token- but not a type-physicalism.

Commitment to physicalism is one tenet CS holds over from behaviorism. Another is the conception of mental properties as functional in character. The difference in this latter respect, as Fodor puts it, is that “whereas behaviorists had permitted only references to stimuli and responses to appear essentially in specification” of the functional relation, his view allows “reference to *other mental states* as well”.³ For a person to be in the mental state of believing that a piece of fruit is edible, for example, is not merely a disposition to ingest the object when sensibly present; it is a disposition that might be functionally dependent upon certain desires (e.g., an urge to satisfy hunger), might function to produce certain expectations (e.g., an

anticipation of satisfaction), etc. The way Fodor's functionalism comes together with token-physicalism is in the view that cognitive states are instantiated according to function in the individual nervous system, and that a given function might be served by different brain mechanisms in different circumstances. The way both come together with the computer model is to view these mechanisms as computational in character.

Thus a central task for the cognitive theorist is to explain how functionally identified cognitive states can interact with one another on a computational basis. One step is to distinguish between propositional attitudes, like beliefs and desires, and other states of the organism that provide their content – internal representations, as Fodor conceives them, that possess "such familiar semantic properties as truth and reference".⁴ It is by virtue of their propositional character that representations play their functional role in cognitive activity, inasmuch as cognition is the processing of propositional content. And it is by forming appropriate attitudes with respect to these propositional contents that the rational organism guides its actions on the basis of experience. In brief, CS accounts for the role of beliefs and other propositional attitudes in our mental life by conceiving them as relationships with other internal states, which in turn possess both (a) semantic properties by which they function in cognitive processes, and (b) physical properties which lend them causal efficacy. The physical properties by which formulae enter into causal relationships are whatever the brain needs for its computations.

The resulting requirement upon the computational model is that some sense must be made clear in which physical computers, of either the mechanical or the biological variety, can deal with meaning-laden formulae in a manner respecting their propositional content. This I take also to be a requirement for making sense of Fodor's remark that "computers are symbol-driven symbol-manipulators", and that "their typical operations consist in the transformation of sets of semantically interpreted formulae", and that "insofar as we view the operations of such machines as computations . . . we are taking these very mechanical processes to be 'endowed with content'".⁵

This is the demand upon the computational model, but at first glance a demand unlikely to be satisfied. For one thing, truth and reference are features involving relationships between formulae and the world at large and hence are not among the physical properties of

an internal formula. Therefore, truth and reference seemingly can play no part in the physical processes of computation. Viewed from a different angle, the problem is that meaning attaches to formulae in computing language only by virtue of programming conventions. But the particular meanings which a programmer assigns to the formulae have no effect themselves upon machine operation. As Fodor himself puts it, "machines typically don't know (or care) what the programs that they run are about; all they know (or care about) is how to run their programs".⁶ But if so, then at least it appears to follow that machines *cannot* operate on meaning-laden formulae.

To rephrase the problem in terms of mental processes: the bearing upon such processes by semantically interpreted formulae is limited by what Fodor calls their "formal"⁷ properties. This predicament he terms the "formality condition", which is tantamount to a sort of "methodological solipsism".⁸ In effect, the predicament is that mental processes "have access only to the formal properties of such representations of the environment as the senses provide. Hence, they have no access to the *semantic* properties of such representations, including the property of being true, of having referents, or, indeed, the property of being representations of *the environment*".⁹ How can this be reconciled with the much emphasized thesis that mental computation involves operations upon meaning-laden formulae?

Fodor's answer is what he describes portentously as "perhaps *the* basic idea of modern cognitive theory".¹⁰ The problem, in his terms, is to devise "some mapping which pairs physical states of the device with formulae in the computing language in such fashion as to preserve desired semantic relations among the formulae".¹¹ The so-called "basic idea of modern cognitive theory" responds to this problem by treating the semantic features of the computing language as corresponding "directly to computationally relevant physical states and operations of the machine", in such a fashion that the "physics of the machine . . . guarantees that the sequence of states and operations it runs through in the course of its computations respect the semantic constraints on formulae in its internal language".¹² What guarantees the correspondence between semantic and causal properties, in lieu of the conventions of natural language, are the "engineering principles" by which the machine operates. In terms of a memorable slogan, the idea is that if the machine is constructed to take care of the syntax, then the semantics can be relied upon to take care of itself.¹³

In summary, CS bills itself as an empirical science, concerned with explaining the causal properties of our propositional attitudes. The causal properties are those captured in conversational language, as in generalizations of the form "seeing that... is a cause of believing that...", and "saying that... is normally caused by believing that...". Beliefs and desires are best described, with respect to their content, in what philosophers call the intentional idiom, but with respect to their causal properties require a physical description. In the first mode the theorist's strategy is "top-down" analysis, yielding intentional descriptions of increasing specificity. In the physical mode his strategy is to invoke the computer model, by which causal interactions are viewed as computational processes. The two approaches converge on the lower levels of analysis, with the specification of subtasks that can be performed computationally by the human nervous system. Since brain processes of this sort are functional in character, CS can be described as a computational functionalism. This means that it views mental states as individuated by function, with different physical operations perhaps involved in different instantiations. Finally, this account of mental processes is bound by the "formality condition", according to which *only* causal properties contribute to the processing of semantic content.

3. THE PROBLEM OF SEMANTIC CONTENT

Not all brain functions, we may assume, are semantically laden, including those of the autonomic nervous system. Let us imagine one such brain function (call it '*A*') in contrast with a brain function (call it '*C*') of the sort to which CS assigns cognitive content. *A* and *C* may be presumed to differ in many respects, among which are differences that account for *C*'s having and *A*'s failing to have semantic content. With reference specifically to *A* and *C*, the problem of semantic content is to specify the features of *C* not shared by *A* that constitute *C*'s having meaning, reference, and so forth. More generally, the problem is to specify the features or set of features by virtue of which a brain state or process is endowed with content, and to explain why those features constitute the possession of content. In its simplest terms, the problem is to provide a scientifically appropriate and satisfactory answer to the question: what is it for a brain state or processes to possess semantic content?

There are responses of various sorts that would be inappropriate. One would be to provide a merely verbal answer such as "for it to refer", or "for it to possess truth value", or simply "for it to have meaning". "Having truth value", "having reference", etc., are different ways of characterizing the possession of semantic content, and what we want to know in pursuing the question is not how else one might characterize the possession of content. What we want to know, rather, is what features constitute the possession of content, meaning, etc.

Another sort of inappropriateness would be illustrated by the claim that semantic content (meaning, reference, etc.) cannot be explained at all. There are various rationales that might be offered for a claim of this sort. One is that of the dualist to the effect that meaning is a form of intention, that intention is a form of consciousness, and that consciousness is intrinsically excluded from scientific explanation. Such a rationale obviously would not be part of CS as characterized above, and there is no reason to discuss it further for present purposes. A second rationale is methodological in character, to the effect that the notion of meaning in its various forms plays a legitimate role in psychological explanation, and does so without itself requiring explanation. There are various ways in which a rationale of this type might be fleshed out. One alternative is to treat the notion of meaning as primitive, which among other things is to accept it both as clear in itself and as capable of contributing to the elucidation of other psychological phenomena. An obvious flaw in this is that the notion of meaning is not clear at all. In fact, it is notoriously ambiguous. There are many conflicting accounts of semantic notions like meaning, truth and reference, developed by philosophers, linguists, and psychologists alike; and for one branch of one of these disciplines to accept as primitive a notion that others consider to be distinctly problematic is not a good tactic for an aspiring science.

Although the practice of some cognitive scientists suggest that the notion of semantic content is being treated as primitive in this fashion, a more reasonable and perhaps more prevalent alternative is that the notion of meaning is sufficiently clear as it stands to play a role in ongoing psychological investigation of other phenomena, but that it is at the same time a notion that ought to be subject to continuing investigation itself. This, at any rate, seems to be the attitude of Fodor, who in a recent publication remarked that the very possibility of

construing beliefs as relationships to mental representations "clearly depends on having an acceptable account of where the semantic properties of the mental representations come from".¹⁴ If CS were to treat semantic content as primitive pure and simple, then of course it cannot provide an adequate account of semantic content itself. The version of CS with which I am concerned in this paper is one which at least considers it an appropriate exercise to attempt to account for semantic content.

There is one further constraint upon the project of providing an account of cognitive content that is dictated by the program of CS itself. Although, in keeping with its physicalist ontology, CS conceives of internal representations as states, processes, or more generally functions of the physical nervous system, it does not encourage any attempt to explain what it is for such representations to have semantic content that is couched primarily in physiological terms. One reason, of course, is its token-physicalism, which precludes any generalized (type-level) characterization of psychological functions in terms of the physical sciences. Another reason, presumably, is that physiology at present possesses neither the experimental nor the theoretical resources necessary to characterize content-laden brain functions in all relevant detail. Yet another is that our understanding of such processes on the psychological level is inadequate to specify what it is about these processes that would have to be characterized in a physiological account of semantic content. All of which is to say just that the problem of semantic content must be solved on a psychological level before we could reasonably attempt a physiological explanation.

The problem of semantic content thus is a problem for psychology specifically, and not one that can be assigned to neurophysiology.

4. HOW COGNITIVE SCIENCE RESPONDS TO THE PROBLEM

CS's answer to the problem of semantic content is composed of two major components. One is a consequence of its token physicalism. If the brain operates on representations in a manner that respects their semantic content, and if both these representations and the operations upon them are physical in nature, then it follows that the brain is capable of physical operations that somehow reflect the semantic features of its constituents. This means that the brain is capable of

running through series of physical procedures that correspond in relevant steps or stages to the procedures by which its representations are processed with respect to their semantic content. This correspondence, of course, must be many-one in character; if it were one-one we would have the makings of a type-physicalism instead.

The other major component is the computer model, in terms of which the nature of these procedures can be conceptualized. Just as a digital computer, through a series of strictly physical transactions, operates upon formulae of its computing language in a manner dictated by their semantic content, so too the brain, as an organic computer, operates on its own internal representations. The result of combining these two components, of course, is Fodor's "basic idea of modern cognitive theory": the notion that there is some mapping which pairs physical states of the system with formulae in its computing language so as to insure that the physical processes of computation will respect and preserve the relevant semantic features of those internal formulae.

CS's answer to the problem of semantic content is a direct extension of this "basic idea". For a brain state or process to have semantic content is for it to be functionally part of the brain's computational operations. More exactly (if we allow that the brain, like other computers, might operate computationally upon some symbols that have no semantic content), to have semantic content is to function as part of the brain's computational routines that operate in accordance with semantic constraints.

It is important to note that for Fodor the computer model is more than merely an heuristic metaphor.¹⁵ As he puts it directly, "what one tries to do in cognitive psychology is to explain the propositional attitudes of the organism by reference to its... computational operations, and... the notion of *computational* operation is being taken literally here".¹⁶ The reason this is important is that the digital computer provides more than a model for conceptualizing the brain's cognitive processes; it also serves as sort of an "existence proof". As Fodor himself recognizes, it seems implausible in the abstract that the "basic idea of cognitive theory" and the "formality condition" should prevail simultaneously – that the brain should operate in a fashion preserving the semantic features of its representations by physical procedures responding only to their nonsemantic properties. Since a properly programmed digital computer allegedly is able to do just

that, it shows how the brain might be likewise capable. Another way of formulating CS's answer to the problem of semantic content is the following: for a brain state or process to have semantic content is for it to function like a formulae with semantic content in the operation of a properly programmed digital computer.

I wish now to show why I think this answer fails, and does so in a manner beyond repair. First I shall argue that systems of the type CS typically relies upon to demonstrate the semantic capabilities of digital computers in fact do not exhibit such capabilities. Next I shall attempt to show that even if such systems (computers programmed to prove theorems) did exhibit semantic capabilities, the operation of such a computer system could not help us understand what it is for a brain state or process to have semantic content. Then I shall suggest reasons for suspecting that CS's unfortunate use of the computer model for this purpose betrays a confusion between two quite disparate senses of information.

5. THE SHORTCOMINGS OF THIS RESPONSE

It is commonplace that digital computers can be programmed to perform certain procedures of logical inference, and logical inference is often thought of as truth-preserving.¹⁷ Since truth is a clear example of a semantic property, it appears an easy matter to program a digital computer so as to insure that the physical properties of its computations preserve a semantic property of at least some of its internal formulae. One way to go about it, briefly, is to devise a canonical notation upon which the machine can operate, to provide for the reduction to canonical form of the logical formulae we wish to deal with, and to provide rules of derivation in terms of computational operations which when correctly followed assure the truth of all formulae derived from true formulae. The key insight of CS is that similar procedures might be developed in connection with other semantic properties¹⁸ – in effect, that by “taking care of the syntax” of representations involving, assertions, memory, perception, etc., the semantic properties of such representations can be made “to take care of themselves”.

The problem with relying on theorem-proving programs as an “existence proof” that semantic features can be dealt with by strictly physical operations is that, in the relevant sense, the algorithms

involved in theorem-proving programs are not *truth*-preserving at all. What they preserve is a formal (physical or configurational) status that might be variously interpreted in varying contexts, but which bears no interpretation whatever as it figures in the algorithm. In one context (practical reasoning), it might be interpreted as moral commitment, in another (epistemic assessment) warranted assertibility, and in yet another of course (deductive proof) it might be truth, but equally well could be mere consistency. Moreover, there is no requirement that the interpretation even have cognitive significance. In the design of electronic switching circuits, for example, it might be relative level of electrical charge. Although the interpretation *truth* is appropriate for certain logical contexts, this interpretation is extraneous to the computational procedure.

There is another sense, to be sure, in which a deductive routine properly executed is truth-preserving. If a properly formulated expression is fed into the routine which happens to have been assigned an interpretation by the programmer or user under which it comes out true, then an expression will be produced as output which under an interpretation consistent with the first will also be true. In brief, true expressions at the input will yield true expressions at the output. But the intermediate states and configurations (including compiling procedures) which the computer goes through in producing the output themselves have no truth value at all. All they have in this regard is one or another formal property which can be interpreted either semantically or nonsemantically at input and output. It is gratuitous to think of such computational procedures as respecting the semantic features of internal formulae on the basis of the physical properties involved in the actual computation. Strictly speaking, the internal formulae have no semantic features to be respected.

Nonetheless, let us grant as a manner of speaking that the internal computational states of a theorem-proving computer have semantic content just when the input formulae from which they derive have been assigned a semantic interpretation by the programmer or user. This still will not help us understand what it is for a brain state to have semantic content. The reason is relatively straightforward. For an internal formula of a computer to have semantic content requires semantic competence on the part of the user, which (under the assumptions of token-physicalism) is for certain brain states of that person to be endowed with semantic content. If the sense in which

brain states of a given human individual possess semantic content is like the sense in which computer states do so, then the former would depend upon another human individual having semantically laden brain states as well, and so on ad infinitum. Thus the sense in which internal states of a digital computer might possess truth and reference cannot help us understand what it is for a human brain state to have semantic content.¹⁹

6. AN INCOHERENT NOTION OF INFORMATION-PROCESSING

Any cognitive psychology “that has a prayer of being true”, Fodor asserts in *The Language of Thought*, “will have to ascribe a special role to the computational states of organism”, i.e., will maintain that “the way that information is stored, computed, accepted, rejected, or otherwise processed by the organism explains its cognitive states and, particularly, its propositional attitudes”.²⁰ Elsewhere in the same book he says that what “cognitive psychologists typically try to do is to characterize the etiology of behavior in terms of a series of transformations of information”.²¹ The connection is that insofar as CS is concerned with computations on formulae with propositional content it is concerned with formulae containing information *that* such-and-such is the case.

Perhaps no term in science has ever been used more variously than the term ‘information’ in cognitivist literature. But there are at least two technical senses on the scene that must be distinguished to avoid confusion, or rather, one technical sense and another semitechnical. The unfortunate assumption that computers in and by themselves are capable of operating with semantically laden formulae, upon which so much weight of modern cognitive theory rests, may stem from a conflation of these two senses.²²

The semitechnical sense is illustrated in the preceding remarks by Fodor, a sense roughly equivalent to propositional content. Information in this sense thus is semantic in character, with meaning and reference and propositional truth value. Let us refer to information of this sort as “info(s)”. One important thing to note about info(s) for present purposes is that it involves a relationship between a symbolic structure and a state of affairs, and that this relationship is representational in character. Another thing to note is that this relationship holds only in a context that establishes a regular correspondence

between the symbol and what it represents. In the case of natural languages generally, the representational relationship between linguistic symbol and state of affairs is established by social contexts of interpretation. In the case of computing languages specifically, the relationship between the physical code sequence entered into the machine and the state of affairs it represents depends also upon the conventions that associate the user's natural language with the particular programming language by which the machine is operated. In the case of brain states, which at very least are not symbols of a natural language, the context of association, by contrast, must be nonconventional in character. Whatever the character of the context, however, it is important to bear in mind that *info(s)* involves a three-term relationship: (1) the meaning-laden symbol (or brain state), (2) the state of affairs or object that it represents, and (3) the context that associates the symbol with the object.²³

The technical sense is that of communication theory, where information (roughly characterized) is reduced uncertainty. Information in this technical sense, which for brevity we may call "*info(t)*", has nothing directly to do with meaning or reference.²⁴ Moreover, the processing of information in this technical sense has nothing to do with propositional content. Although there is a sense in which a signal at the output of a communication channel conveys information about what occurred at the input, the term 'about' in this sense means "with respect to" or "relative to". In particular, the output does not provide *info(s)* that such-and-such occurred at the input, since (among other reasons) *info(t)* abstracts entirely from interpretive contexts in which the symbol-occurrences across the channel have propositional meaning.²⁵ In the only sense in which *info(t)* even approximates the "aboutness" of semantic content, it involves only a two term relation between input and output occurrences.

Now computers clearly are information-processing systems, in the sense specifically of *info(t)*. Communication theory was developed originally for application to electronic circuitry, and has remained an important tool in the design of digital computers. It should be equally clear, on the other hand, that computers as such (i.e., as uninterpreted symbol-manipulators) are incapable of dealing with *info(s)*. Although they can compute changes on symbol-strings that happen to have been given this or that interpretation by a human user, these interpretations do not enter into the computational process. Nor is it relevant to the

computational process whether the symbol-strings are interpreted at all, or whether if interpreted their interpretations are semantical in character. This is precisely the point of Fodor's "formality condition". And although to be sure computers are generally constructed so that the results of their computations sustain the interpretations intended by their designers and users (a paraphrase of "the basic idea of modern cognitive theory"), this in itself does not help us understand what it is for the human mind to deal with semantic content.²⁶

To put it another way, the computer model is powerless as an aid to genuine understanding of what semantic content amounts to and how it originates.²⁷ Contrary to the bemused talk of some cognitive scientists, computers are not info(s)-processing systems. To adopt a well-known image from another context,²⁸ computers are semantic juice-extractors which themselves are devoid of semantical juice.

Insofar as CS is responsible for a coherent account of what it is for the human brain to operate with semantic content, Fodor is being unduly optimistic in attributing to it so much as a "ghost of a chance" in its current form, relying as it does on the digital computer as a model of information-processing. For while the digital computer is a superlative processor of info(*t*), it has no capabilities at all with info(s).

To have a "ghost of a chance" of explaining how the brain operates with semantic content, a research program should start with something we already understand, and work toward an account of what needs explaining. As matters stand, we don't understand what it is for symbols to be endowed with content, nor how they come to be so endowed. To assume that we do understand these things, or even that we understand them well enough to get a theory of cognition underway, is simply a bad research strategy for a fledgling science. How the mind operates with semantic content is scarcely understood at all; rather, it is what requires explaining. One thing we do understand quite well at this point, however, is how computers operate with info(*t*). And very likely the human brain operates with info(*t*) in much the same manner. A really promising research strategy, accordingly, would be to begin with the basic features of info(*t*)-processing, illuminated by the results of communication theory,²⁹ and to work toward an account of how info(*t*)-processing systems might acquire the capability of operating with info(s).

A crucial part of such a theory would be an account of the sense in which brain states or processes might *represent* salient features of the

perceptual environment. Relying on the resources of communication theory, I wish in the final section to indicate in a very general fashion one form such an account might possibly take.

7. OUTLINES OF A COMMUNICATION-THEORETIC ACCOUNT OF REPRESENTATION

What is it for a brain state or process C to *refer* to a certain aspect of the organism's environment, to have a certain status (e.g., *truth* or *falsehood*) which depends upon the environment's being in such-and-such a state, or in any other fashion to have *meaning* (e.g., denotation) that picks out certain objects to the exclusion of others?

Clearly, it is for C to bear a certain kind of relationship to the object it denotes, the state of affairs to which it refers, etc. But what kind of relationship? It is tempting to rely on reassuring labels – to say, for example, that it is a representational relationship. But apart from how we label it, what does the relationship amount to? In the spirit of what is best in contemporary cognitivism, let us attempt an answer in functional terms.

The relationship is between brain state C and a state or aspect of the organism's environment, which we may designate E_c . The location of C in the brain, its influence upon the organism's efferent nervous system, its interaction with other sectors of the afferent system, and undoubtedly many other of its specific features, will depend upon whether E_c is perceptually present or absent, familiar or unfamiliar, feared or desired. Such matters also will depend upon whether C is a percept, a concept, a belief or a memory, or a plan for the guidance of future behavior. All such factors, and many more presumably, would require separate treatment in any adequate account of the mind's cognitive functions. It is painfully obvious that anything to be said about C and its relationship to E_c that abstracts from such considerations will have to be very schematic and general indeed. But some things *can* be said about the function of C in general, whether as percept or concept or memory image, and they can be said in terms of communication theory.

For one thing, this relationship between C and E_c enables the organism to discriminate E_c from other states of the environment, and if desired to direct its activities towards E_c specifically. In familiar terms, the relationship enables the organism to direct its attention

toward E_c in distinction from other objects, circumstances, or states of affairs. This, on a basic functional level, is what is involved in C 's being a *representation* of E_c , and at the same time what it is for E_c to be the object of the organism's *intentional* awareness.

A relationship of this general character is present between C and E_c just in case the pair of occurrence-sets (E , E_c) possesses a sufficiently high degree of mutual info(t).

For this formulation to be useful, it must be accompanied by definitions of 'occurrence-set' and 'mutual info(t)', and an explanation of sufficiency in degree of mutual info(t).

Any object or state of affairs E_c can be characterized with respect to the conditions or events the joint occurrence of which constitutes its obtaining or taking place. Furthermore, a state of affairs can be so characterized on different levels of specificity. My pen's being on my desk, for example, may be characterized simply as the placement of this object on this surface, or as the appearance of a black cylinder upon the flat green expanse in front of me, or as the collocation of these several pen-parts at a certain location in space, etc. The relevant level of characterization depends upon the organism's involvement with E_c , and may be expected to vary as the circumstances of that involvement vary. In any case, the arrangement or disposition of objects or features that constitute a given state of affairs will be identifiable as one among several alternative arrangements, any one (but only one at a time) of which the organism may encounter under relevant circumstances. And, within rough limits of accuracy, each alternative can be assigned a numerical probability of occurrence. For a given state of affairs E_c , the *occurrence-set- E_c* is the set consisting of E_c and all relevant alternatives with nonzero probabilities of occurrence. The membership of this set varies with the circumstances, but includes at least E_c and one relevant alternative (perhaps just the absence of E_c itself).

Any brain state or process C can be characterized (assuming an adequate neurophysiology) with respect to the activity or inactivity of the nerve cells or components of nerve cells by which C is constituted. Such a characterization also could cite features at various levels of specificity, including interactions and sequences of the nerve-process involved. For any given arrangement of neuronal events thus characterized, there will be a set of alternative arrangements each of which will have some distinct effect upon the organism's behavioral or

perceptual activities. Within rough bounds, each relevant alternative can be assigned a numerical probability of occurrence. For a given brain state C , the occurrence-set- C is the set of all relevant alternatives with a nonzero probability of occurrence.

Details at this point are unimportant, and would only be conjectural at our present stage of knowledge. What is important is to be persuaded that both E_c and C are single structures within a set of relevant alternatives, that in either case these alternatives consist of different structures of basically the same components, and that the representation of E_c by C consists in a particular relationship between these two sets of alternatives. This relationship can be explicated in terms of the concept of mutual information, a technical concept from communication theory.

To bring the resources of communication theory to bear, we may conceive E_c and C as input and output respectively of a communication channel, with specific capacities for $\text{info}(t)$ -transmission. One capacity is its ability to indicate individually at its input a certain range of alternative occurrences, specified in terms of the *entropy* of the set E_c .³⁰ The entropy of E_c is the average amount of $\text{info}(t)$ carried by the occurrence of members of that set, where the $\text{info}(t)$ carried by an individual member is equal numerically to the number of times the *a priori* probability (probability prior to occurrence) of that event must be doubled to equal unity (its probability after occurrence).³¹ Another capacity is the channel's ability to convey $\text{info}(t)$ from input to output with a certain degree of fidelity, i.e., without equivocation. Equivocation is the average amount of $\text{info}(t)$ lost per signal across a channel, defined technically as the average uncertainty left regarding the input-occurrence after a signal is received at the output.

The mutual $\text{info}(t)$ of a channel's input with respect to its output is the difference between the entropy of its input and the equivocation of its input with respect to its output. Mutual $\text{info}(t)$ thus is a measure of the channel's capacity for reliable $\text{info}(t)$ transmission: the average amount of $\text{info}(t)$ it can receive at its input, minus the average loss of information in the course of transmission. The mutual $\text{info}(t)$ of the channel $E_c - C$ characterizes the capacity of this channel as a reliable conveyor of $\text{info}(t)$ from E_c to C .

To characterize the channel $E_c - C$ in terms of its $\text{info}(t)$ -processing capacities is not to say anything in particular about its physical

features. Although for any specific channel $E_c - C$ presumably there is a specific series of causal interactions establishing the connection between input and output, the physical nature of this series does not figure in its description as an $\text{info}(t)$ channel. Insofar as the representational relationship between C and R_c consists in there being a sufficiently high degree of mutual $\text{info}(t)$ between these two occurrence-sets, accordingly, the particular physical connection between C and E_c is not an essential component of that relationship. This is as it should be with semantic representations.³²

It is not sufficient for C to function as a representation of E_c merely that the occurrence-sets C and E_c stand in a relationship of mutual $\text{info}(t)$. Any two sets of events that are not statistically independent possess *some* degree of mutual $\text{info}(t)$. For C to function as a representation of E_c requires that occurrence-sets C and E_c possess a degree of mutual $\text{info}(t)$ that is sufficient to guide the organism's activity with respect to E_c , as distinct from other environmental circumstances.

To indicate roughly how this notion of "sufficient for guidance" is to be fleshed out, it should be helpful to note a second thing that can be said about the function of C in general, abstracting from its role as percept, concept, memory image, et al. Inasmuch as C guides the organism in discriminating selected aspects of the environment, and in directing its activities toward those aspects specifically, it must be capable of exercising this guidance in a manner that varies with environmental circumstance. If E_c becomes more elusive in the organism's interaction with it (takes evasive action if E_c is an animal under pursuit; becomes confused with other circumstances if E_c is an object of recall; etc.), then C itself may be called upon to deliver more $\text{info}(t)$ about E_c than was previously present. One way this might occur is for the organism to improve its vantage point with respect to E_c (approaching it more closely; finding a better perspective; etc.), thus increasing the fidelity of the channel $E_c - C$. Another way would be for C to alter in such a fashion as to respond to a wider range of component-occurrences on the part of E_c (to indicate more or different details of the environmental circumstances in question), thus increasing the amount of $\text{info}(t)$ that can be entered into the channel. In either case, responses on the part of the behaving organism alter the communication channel $E_c - C$ in a fashion that increases its mutual $\text{info}(t)$.

A third thing that can be said about the function of C in general is that, whatever specifically its role in the guidance of the organism's mental or physical activity, it will perform that role with a minimum expenditure of the nervous system's $\text{info}(t)$ -processing resources. Although it is far from clear how exactly the brain's $\text{info}(t)$ -processing resources ought to be measured, it is clear that those resources are not unlimited.³³ And of the limited resources available, considerable quantities undoubtedly are dedicated to functions that have nothing to do with the processing of $\text{info}(s)$. Since, moreover, the human central nervous system must have been under considerable pressure in the development of the species to respond selectively to an increasing range of different environmental circumstances, it presumably has become a matter of basic importance to conserve the amount of $\text{info}(t)$ -processing capacity expended by the nervous system in any given cognitive task. In general, one may reasonably conjecture, the nervous system commits as little of its $\text{info}(t)$ -processing capacity to a given task of thought, perception, memory, etc., as is compatible with its successful completion.

Metaphorically, C can be thought of as a system of neuronal activities that "locks on" to those features of E_c to which the organism directs its selective attention, or as a receiver that "resonates" to the signals emitted by E_c 's occurrence. However conceptualized, the association between C and E_c functions to maintain a working balance between two opposing requirements. On one side is the requirement that the communication channel $E_c - C$ maintain a level of mutual $\text{info}(t)$ sufficiently high to enable the organism to fix its attention upon or otherwise to operate with respect to E_c specifically, as distinct from other features of the organism's environment. On the other side is the requirement that C function in this connection with minimum expenditure of the organism's overall $\text{info}(t)$ -processing capacities.

Brain state C functions as a *representation* of environmental state E_c when together with E_c it constitutes a communication channel governed by these two constraints – the interacting requirements of faithful $\text{info}(t)$ -transmission from environment to nervous system, and of efficient information(t)-processing on the part of the latter. Under these circumstances, E_c constitutes what in this discussion we have been referring to as the *semantic content* of brain state C . The dual set of constraints governing the association between these two structures constitute the context by which C is endowed with $\text{info}(s)$ -content –

the third term of the semantic relationship by which $\text{info}(t)$ is converted to $\text{info}(s)$.

8. RECAPITULATION AND PROGNOSIS

The preceding remarks are too schematic, and much too scanty, to count as a theory of semantic content. At best, they constitute the basic theme of such a theory, a theme more fully worked out in Sayre (1976). Much work remains to be done before this basic theme could begin to take the form of a substantive program for inquiry into the brain's cognitive capacities.

Nonetheless, this basic approach does show a promise of yielding a research program which has something more than merely a "ghost of a chance" of succeeding. To have an appreciable chance of succeeding, a research approach should begin with something we really understand. One thing we do not understand at all is what it is for brain states to have semantic content. For the resolution of this problem it is no help to fall back upon the computer model, for computers have no competence with $\text{info}(s)$. Hence the currently fashionable reliance of CS upon the computer model stands no chance of helping us understand the nature of semantic content in the human information-processing system.

A more promising approach would start with something we understand already, and work toward an account of what needs explaining. What needs explaining is how the brain deals with $\text{info}(s)$. One thing we already understand is how symbol-manipulation systems deal with $\text{info}(t)$. Both brains and computers operate with $\text{info}(t)$ – that much seems beyond reasonable doubt – and very likely they do so in similar ways. A really promising approach, accordingly, would be to begin with basic capacities for $\text{info}(t)$ -processing that appear to be within the competence of mechanical and biological symbol manipulation systems indifferently, and to work toward an account of how biological systems acquire the additional competence to deal with $\text{info}(s)$. In the process, we might expect to find considerable advantage in reflecting on how digital computers deal with $\text{info}(t)$, and hence will find continued use for the computer model. Also, we undoubtedly will find that the modes in which the human nervous system processes $\text{info}(t)$ involve functional interactions of various sorts, so the functional approach will remain in business. I have attempted to indicate in this

paper one direction an approach of this more promising sort might possibly take.

One aspect of this proposed approach that some cognitive scientists will find unattractive is that it relies implicitly upon knowledge of how the brain functions in its $info(t)$ -processing tasks. Thus CS, if it were to follow such an approach, could not hope to remain autonomous from the biological sciences. Whether this should render the approach unacceptable is a matter for debate on a methodological level. For my part, it seems clear that it should not.

NOTES

¹ Although the notion of intentionality finds various characterizations in recent cognitivist literature, the core sense is that of being *about* something or having a *meaning*. Particularly helpful discussions of the "top-down" mode of analysis can be found in Haugeland (1978) and Dennett (1978, ch. 7).

² See Fodor (1981b, pp. 18-19); also Pylyshyn (1981, p. 159).

³ Fodor (1981b, p. 10), author's emphasis.

⁴ Fodor (1975, p. 32).

⁵ Fodor (1981b, p. 23), author's emphasis.

⁶ Fodor (1981b, p. 207).

⁷ Fodor (1981b, p. 231). In this use, 'formal' is roughly equivalent to 'physical' or 'causal', and more closely equivalent to 'syntactic' in the broad sense of nonsemantic.

⁸ *Ibid.*

⁹ *Ibid.*, author's emphasis.

¹⁰ Fodor (1981b, p. 240), author's emphasis.

¹¹ Fodor (1975, p. 73).

¹² This and the quoted expression following are from Fodor (1975, p. 66).

¹³ From Haugeland (1981, p. 23).

¹⁴ Fodor (1981a, p. 123).

¹⁵ In (1981b, pp. 100-123, and elsewhere), Fodor is explicit in opposing his position on this matter with that of Dennett, who has maintained in print that thinking of mental processes as operations on semantically characterized symbols has heuristic but no literally descriptive value.

¹⁶ Fodor (1975, p. 76), author's emphasis.

¹⁷ A clear statement of logical inference as a manner of dealing with semantic properties that can be easily automated may be found in the Introduction of Haugeland (1981b, pp. 22-23).

¹⁸ As Fodor (1981b, p. 221) puts it, the insight is that we might be able to formulate a "canonical representation of a sentence... which captures not just its logical form... but which also provides an appropriate domain for principles which govern... memory storage, or interaction with perceptual information, or learning, or whatever".

¹⁹ This argument applies only to what Haugeland (1981) calls "automatic formal

systems" (p. 10), which receive formal inputs exclusively (not counting wiring changes as input). The story might be different with computing systems that interact nonsymbolically with their environment, and in that fashion develop semantic competence by processes paralleling those of human-language evolution.

²⁰ Fodor (1975, p. 75).

²¹ Fodor (1975, p. 52).

²² Early cognitivist literature clearly suffered from this confusion. For discussion of examples, see the Preface of Sayre (1976).

²³ A similar point is made in Ringle (1981, p. 45).

²⁴ This has been stressed repeatedly by communication theorists, beginning with Hartley (1928, p. 538), and Shannon (1949, p. 3). To insist upon this is not to ignore the various attempts that have been made subsequently to develop an account of *info(s)* on the basis of communication theory, as presaged in Weaver (1949, p. 116).

²⁵ If the output of a communication channel provides a noiseless indication of a particular input event, and if someone *knows that* the channel is noiseless in this respect, then the output provides *info(s)* that the input event in question occurred. Intentionality of *knowing that* begets intentionality of *being information that*; the latter does not spring full-blown from any amount of *info(t)*.

²⁶ Stich apparently believes that any problem of this sort has been solved by Dennett. In a comment on 'Artificial Intelligence as Philosophy and as Psychology' (Dennett, 1978), Stich remarks that psychology once was haunted by the argument (as he says, roughly) that internal representations "serve no explanatory function unless they can be interpreted: but interpretation requires an interpreter, an homunculus; and relying on homunculi is either circular (since they have just the abilities we are trying to explain) or leads to an infinite regress". But, he goes on to say, "Dennett scuttles the argument by showing how the use of progressively stupider homunculi can avoid both circularity and infinite regress" (privately circulated communication). The way Dennett is supposed to have shown this is by citing the technique of "top-down" analysis into "nested intentional systems" with each successive level describing a less demanding function, to be performed, as it were, by progressively stupider homunculi. Eventually, Dennett says, this "nesting of boxes within boxes lands you with homunculi so stupid . . . that they can be, as one says, 'replaced by a machine'. One *discharges* fancy homunculi from one's scheme by organizing armies of such idiots to do the work" (Dennett, 1978, p. 124), author's emphasis; Dennett might have noted the similarity to Selfridge's 1959 Pandemonium program. Whatever the advantages for AI of this program of "reduction by attrition," it seems to offer no help in the cognitivist's critical problem of explicating the nature of semantic content. For one thing, it is not clear that there is any intelligible sense to the notion of breaking the "functions" of meaning and reference into "more and more stupid subfunctions". For another, it should be noted that the problem of semantic content as discussed in the text above rises precisely on the most highly analyzed level—that of canonical representation—at which there is no "simpler" task of interpretation to be discharged by "stupider" homunculi.

²⁷ Note that it would be fruitless to augment the computer model with the notion of a built-in semantic interpreter—in effect, to take the computer-cum-programmer as the basic model. For the upshot of applying such a model to mental activity would be explicitly to postulate a "ghost in the machine". See Heil (1981) for a similar point. The

proscription against ghostly homunculi, to which cognitive scientists generally pay lip service at least, should be accompanied by a more urgently needed proscription against attributing ghostly "homunculus functions" to the machine, specifically a proscription against intentionally described machine functions. In the contemporary climate of cognitivist euphoria, the latter is needed to preserve the spirit of the former.

²⁸ Hempel (1945).

²⁹ Dretske (1981) appears to be an attempt of this general sort. The reasons I think Dretske's attempt does not succeed are (1) that his account is not in terms of $info(i)$ at all, but rather relies upon an embellishment that runs counter to technical communication theory in several basic respects, and (2) that his account of the nature of $info(s)$ involves several assumptions about causal relationships that may not hold good in the actual world. These criticisms are detailed in my commentary on Dretske's book in *The Behavioral and Brain Sciences* 6(1), 1983.

³⁰ The relationship between this entropy and that of thermodynamics is discussed in Sayre (1976, Chap. 3).

³¹ Mathematical definitions of entropy, equivocation, and mutual information may be found in Sayre (1976, Chap. 2, *et al*). The definitions in question are part of the standard communication theory, not embellished as in Dretske (1981).

³² Smoke is a natural indicator of fire because it is naturally caused by fire. For the most part, however meaningful symbols are not caused by the things they represent. Similarly, it is not the *causal* aspects of the process by which brain state C is produced that make it a representation of environmental state E .

³³ For citations of estimates see Sayre (1976, p. 160, footnote 6).

REFERENCES

- Block, Ned (ed.): 1981, *Imagery*, The MIT Press, Cambridge.
- Dennett, D. C.: 1978, *Brainstorms*, Bradford Books, Montgomery, Vermont.
- Dretske, Fred I.: 1981, *Knowledge and the Flow of Information*, The MIT Press, Cambridge.
- Dreyfus, Hubert L.: 1981, 'From Micro-Worlds to Knowledge Representation: AI at an Impasse', in Haugeland.
- Fodor, Jerry A.: 1975, *The Language of Thought*, Thomas Y. Crowell Company, New York.
- Fodor, Jerry A.: 1981a, 'The Mind-Body Problem', *Scientific American* 214, 114-23.
- Fodor, Jerry A.: 1981b, *Representations*, Bradford Books, Montgomery, Vermont.
- Hartley, R. V. L.: 1928, 'Transmissions of Information', *Bell System Technical Journal* 7, 535-63.
- Haugeland, John (ed.): 1981, *Mind Design*, Bradford Books, Montgomery, Vermont.
- Haugeland, John: 1978, 'The Nature and Plausibility of Cognitivism', *The Behavioral and Brain Sciences* 1, 215-226. Reprinted in Haugeland (ed.), (1981).
- Haugeland, John: 1982, 'Weak Supervenience', *American Philosophical Quarterly* 19(1), 93-103.
- Heil, John: 1981, 'Does Cognitive Psychology Rest on a Mistake?', *Mind* 60, 321-42.
- Hempel, Carl: 1945, 'On the Nature of Mathematical Truth', reprinted in Feigl and

- Sellars (eds.), *Readings in Philosophical Analysis*, Appleton-Century-Crofts, Inc., 1949.
- Pylyshyn, Zenon: 1981, 'The Imagery Debate: Analog Media versus Tacit Knowledge', in Block.
- Ringle, Martin: 1981, 'Artificial Intelligence and Semantic Theory', in T. Simon and R. Scholes (eds.), *Language, Mind and Brain*, LEA, Inc., Hillsdale, N.J.
- Sayre, Kenneth M.: 1976, *Cybernetics and the Philosophy of Mind*, Rutledge and Kegan Paul.
- Searle, John R.: 1981, 'Minds, Brains, and Programs', in Haugeland.
- Shannon, Claude E.: 1949, *The Mathematical Theory of Communication*, The University of Illinois Press, Urbana, Illinois.
- Toulmin, Stephen: 1953, *The Philosophy of Science*, Hutchinson University Library, London.
- Weaver, Warren: 1949, 'Recent Contributions to the Mathematical Theory of Communication', in Shannon.

Department of Philosophy
University of Notre Dame
South Bend, IN 46556
U.S.A.

PART IV

MENTALITY AND INTENTIONALITY

THE PRIMACY OF THE INTENTIONAL

1. INTRODUCTION

According to the thesis of the primacy of the intentional, the reference of language is to be explicated in terms of the intentionality of *thought*. The word "*Pferd*," for example, refers to horses in so far as it is used to express thoughts that are directed upon horses. But most contemporary philosophers of language, until recently at least, have held that the intentionality of thought is to be explicated in terms of the reference of *language*. But no such explication is at hand.

I shall suggest what one can say about the reference of language if one presupposes the primacy of the intentional. I will not use any undefined semantical concepts, since all such concepts, I believe, can be explicated only by reference to intentional concepts. Therefore I will not speak of 'inner systems of representation', 'inner speech acts', or 'inner language'. But I shall make use of a number of *intentional* concepts. In terms of these and certain other familiar concepts, I will formulate intentional definitions of such semantic concepts as *sense* and *reference*. I will try to show how this approach will throw light upon a number of philosophical questions (e.g., 'How are we to interpret the "he, himself" locution?' and 'Do demonstratives and proper names have senses?').

LANGUAGE INTENTIONALLY CONSIDERED

Two assumptions underlie what might be called the orthodox approach to the analysis of sense and reference. One of these is ontological and the other is psychological. Each seems to me to be questionable.

(1) The first assumption is this: for each use of any well-formed indicative sentence in our language, there is a *proposition* which is the meaning of that sentence in that use.

I would say that this is not an assumption with which we should *begin* our investigations. It is, at best, a conclusion we might draw at the *end* of our investigations. For the 'propositions' that are thus presupposed will not be restricted to the abstract objects now commonly called 'states of affairs'.

Those 'singular propositions' that would constitute the meanings of sentences containing demonstratives and proper names would be contingent things, dependent for their being upon those individual things that are thought to 'enter into them'. And these singular propositions are strange entities.

There is a certain plausibility in the assumption that there are what might be called 'concrete states or events'. It is reasonable to assume, after all, that if a contingent thing x has a certain property P , then there is also that concrete state or event which is x having P . But concrete states or events will not do the work that is required of singular propositions. For *false* or *nonoccurrent* singular propositions would be needed to constitute the meanings of sentences that are false. If you believe that there are twelve planets then, even if there are only nine planets, the singular proposition theory requires the existence of that contingent thing which is *there being twelve planets* (or *that there are twelve planets*).

(2) Since, presumably, we can believe only what we can *grasp*, or *conceive*, the singular proposition theory presupposes further that the believer is able to grasp or conceive, not only those abstract objects which are properties and states of affairs, but also those contingent 'singular propositions' constituting the meanings of sentences containing demonstratives and proper names. It is presupposed that demonstrative terms and proper names, like such definite descriptions as 'the tallest spy', have what might be called a *Fregean sense*. A Fregean sense is a property which is such that a term having that sense designates a thing only if the thing has that property. One then looks in vain for those properties constituting the Fregean senses of demonstratives ('this', 'I', 'you', 'now') and of proper names ('Tom', 'Cicero').

The conception of intentionality that I shall set forth is considerably simpler than the foregoing and it has a number of advantages as a basis for understanding language.

DE RE BELIEF

I will assume that intentional attitudes involve a relation between a person and a *property* or *attribute* rather than between a person and a *proposition*. I shall set forth this view in some detail in application to occurrent belief – or *believing* – and somewhat more sketchily in connection with *endeavoring* and *perceiving*. What is said about these attitudes may also be applied, *mutatis mutandis*, to *thinking*, or *considering*.¹

We begin, then, with *believing*.

I presuppose that believing is essentially a matter of believing certain properties '*directly of oneself*'. (We could also say: 'attributing certain properties directly to oneself'.) The fundamental doxastic concept may be expressed by the locution

x believes directly of y the property of being-F

a locution that is taken to imply

x is identical with y.

Direct belief, that is, is belief about oneself. The letter 'F' in 'being-F' is schematic and may be replaced by any English predicate – say, 'wise'. The result of such replacement will be the English *term*, 'being-wise' – a term designating a property.

What would it be, then, for one to have a belief about something *other* than oneself? Suppose, for example, that I believe *you* to be wise. In this case, there is a property I believe directly of me; this property relates me just to you; and because of this relation, I can be said to believe being-wise *indirectly* of you. Such a belief might come about under the following circumstances. You are the person I am talking with; and the property of talking with just one person and with a person who is wise is a property that I attribute directly to myself.

We may be tempted, then, to characterize indirect attribution this way:

x believes being-F indirectly of y =_{Df} There is a relation *R* such that

- (a) *x bears R only to y and*
- (b) *the property of bearing R to just one thing and to a thing that is F is one that x believes directly of x.*

Then we could produce a general analysis of *de re* belief by saying that *x believes y to be F* provided only that *x* believes being-*F* either directly or indirectly of *y*. But this preliminary account may be more latitudinarian than it should be.

Consider the following example, adapted from one suggested by Keith Donellan.²

- (1) I believe that the person I am looking at is a member of the Temperance Union;
- (2) I believe, mistakenly, that the person I am looking at is the only one at the party who is drinking a martini;

- (3) putting two and two together, I judge that the only one at the party who is drinking a martini is a member of the Temperance Union; and
- (4) as it happens, the only one at the party who is drinking a martini is, not the person I am looking at, but the person who is standing behind me.

Application of our formula above to this example would require us to say, not only that the person I'm looking at, but also that the person who is standing behind me, is such that I believe that *he* is a member of the Temperance Union. My belief, so to speak, 'points to' both individuals. But if it is a belief that is *directed upon* one of them, then it is not also a belief that is directed upon the other.

If my belief is directed upon a specific individual, say, *you*, then I have some conception of just *who* you are. But in what sense of 'conceiving who you are'?

I will first put the answer somewhat loosely. Then I will attempt a more precise formulation.

If my belief is directed upon you, then there are two things to be said about the *justification* that I have for that belief: (1) the property by means of which I relate myself uniquely to you is one such that I am justified in believing that there is just one thing having it; and (2) at the moment I have no *better* way of identifying you.

In putting the account of indirect attribution more exactly, I will use '*x* believes himself to be *F*' as short for 'the property of being *F* is such that *x* believes it directly of *x*':

D1 *x* believes indirectly with respect to *y* that it is *F* =_{Df} There is a relation *R* such that:

- (1) *x* bears *R* just to *y*;
- (2) *x* judges himself to bear *R* to just one thing and to a thing that is *F*;
- (3) *x* is more justified in judging himself to bear *R* to just one thing than in not judging himself to bear *R* to just one thing; and
- (4) if *x* judges himself to bear *R* to the thing he bears *S* to, then he is at least as justified in judging himself to bear *R* to just one thing as he is in judging himself to bear *S* to just one thing.

If, in judging that he is talking with a member of the Temperance Union, *x* has a belief that is directed upon *y*, then: (1) *x* is more justified in believing

that there is one and only person he is talking with than in not having such a belief; and (2) if x also judges that the person he is talking with is, say, *the Chairman* of the Temperance Union, then he, x , is at least as justified in believing that there is just one person he is talking with as in believing that there is just one person who is the Chairman of the Temperance Union.

And now we may set forth our definition of *de re* belief:

D2 x believes with respect to y that it is $F =_{\text{Df}}$ The property of being F is such that x believes it either directly or indirectly of y .

On this account, the *de dicto* locution, ‘ x accepts the proposition that p ’ could be defined as: ‘There is a y such that x believes with respect to y that it is true’.

We should note, in passing, that this epistemic conception of *de re* belief may also be adapted to the propositional conception of intentionality. One could say:

x believes with respect to y that it is $F =_{\text{Df}}$

- (1) x believes that the G is F ;
- (2) y is the G ;
- (3) x is more justified in believing that there is just one thing that is G than in not believing that there is just one thing that is G ; and
- (4) if x believes that the G is the H , then x is at least as justified in believing that there is just one thing that is G as he is in believing that there is just one thing that is H .

OBJECT AND CONTENT

The distinction between *object* and *content* is essential to understanding language. In saying something to you, my concern may be to get you to *believe* something. Or it may be to get you to *do* something. Or it may be merely to get you to *think of* something. In each case, there is a distinction between object and content. Thus there is the object, or there are the objects, that I want you to believe something *about*, or to do something *to*, or to think of in a certain *way*. And there is what it is that I want you to believe about the object, or what it is I want you to do to the object, or what it is I want you to think of the object as being.

Suppose you are driving a car in which I am a passenger and I say to you urgently: ‘That green Chevrolet over there is out of control!’ If you reply, ‘That’s not a Chevrolet,’ you may correct what I say, but you will have

misinterpreted my message. I might reply, if there is time, 'But whatever make of car it is, it's out of control!'³ The words 'That green Chevrolet over there' in my original utterance did not express any part of the *content* I meant to convey. I used them only so that you would pick out the *object* I want to convey something about. The *content* of my message was expressed by the words 'out of control'.

Since, in believing, one directly attributes a property to oneself and in so doing one may indirectly attribute another property to some other thing, we may distinguish between the *direct* and the *indirect* object and content of belief. And analogously for the other intentional attitudes. Thus we may distinguish between the object and the content of *thinking of* and between the object and content of *endeavor*.

PERCEIVING AND ENDEAVORING

The concept of *meaning to convey* which is central to understanding the intentionality of language presupposes the further intentional concepts of *perceiving* and *endeavoring*.

Perceiving is a paradigm case of what we have called 'indirectly believing', or 'indirect attribution'. In perceiving the car, say, to be moving, one directly attributes to oneself the property of being appeared to by just one thing and to a thing that is a car and is moving. The intentional element in perceiving is this:

- D3 x perceptually takes y to be $F =_{\text{Df}}$ y and only y appears G to x ; and x directly believes of x the property of being appeared G to by just one thing and by a thing that is F .

An adequate analysis of perceiving would require an epistemic concept, one implying the justification of belief. Thus we might say:

- D4 x perceives y to be $F =_{\text{Df}}$ x perceptually takes y to be F ; and x is justified in believing that y is F .

The intentional element in the concept of endeavor may be expressed by 'The property of being F is such that x endeavors to have it'. In some contexts, 'acts with the intention of' will be used instead of 'endeavors to'. *Indirect* endeavor is analogous to indirect belief. (Once again, I will use ' x believes himself to be F ' as short for ' x believes being- F directly of x '.)

- D5 x indirectly endeavors that y be $F =_{\text{Df}}$ There is a relation R such

that:

- (1) x will bear R just to y ;
- (2) the property of bearing R to just one thing and to a thing that is F is such that x endeavors to have it;
- (3) x is more justified in believing himself to be such that he will bear R to just one thing than in not believing himself to be such that he will bear R to just one thing; and
- (4) if x believes himself to be such that the thing he will bear R to is the thing he will bear S to, then he is at least as justified in believing himself to be such that he will bear R to just one thing as he is in believing himself to be such that he will bear S to just one thing.

Our definition of *de re* endeavor is now this:

- D6 x endeavors that y be F =_{Df} Either
- (a) x is identical with y and endeavors to have the property of being- F ; or
 - (b) x indirectly endeavors that y be F .

We need also the concept of endeavoring to do one thing *for the purpose* of bringing about another thing (or *in order to* bring about another thing). For present purposes this simplification will do:

- D7 x endeavors to bring about P and does so *for the purpose* of bringing about Q =_{Df} x endeavors to bring it about that his endeavor to bring about P cause Q .

This concept thus presupposes the concept of causation.

MEANING TO CONVEY

The relation between thought and language may now be described by reference to the concept of *meaning to convey*. Meaning to convey is *endeavoring to convey*. What, then, is *conveying*? If I want to convey something to you, then I have a certain thought I want to communicate to you; and this means, in part, that there is something I want to cause you to think of. (I here take 'cause to think of' broadly – to cover both the case where one is caused to *begin* to think of a certain thing and also the case where one is caused to *continue* to think of that thing.)

But conveying is more than merely causing to think of. Let us consider two

cases of causing to think of that are not cases of conveying.

- (1) I inject you with a certain drug that makes people paranoid and then I present you to Mr. Jones. The result is that you believe that Mr. Jones desires to persecute you. But even if it had been my intention to cause you to believe this, we cannot say, taking 'convey' in its present sense, that I have *conveyed* this to you.
- (2) Kant cites the following case of intended deception that is not a case of lying: 'I may wish people to think that I am off on a journey, and so I pack my luggage; people draw the conclusion I want them to draw... I have not lied to them, for I have not stated that I am expressing my opinion'.⁴

Kant's remark may suggest that, in order to be able to tell you anything, I must first tell you that I am going to tell you something, and in order to be able to convey anything, I must first convey that I'm going to convey something. This type of regress would hardly be acceptable.

What, then, does *conveying* involve that mere *causing to think of* does not involve? I suggest that there are three marks that are characteristic of conveying but need not hold of mere causing to think of.

One mark of conveying may be illustrated by this: If I convey something to you, I do so by causing you to believe that *I* am thinking of that something. This is not what happens when, merely by injecting a drug, I cause you to have a certain belief.

Secondly, if I convey something to you, then my *purpose* in causing you to believe that I am thinking of the thing in question is that of causing *you* to think of that thing.

And a *third* mark of conveying pertains to the fact that one is *addressing* someone. If I mean to convey to you the thought that so-and-so, then I intend to cause you to believe that *I intend* to cause you to have the thought that so-and-so. Hence conveying presupposes a complex belief situation. We must be able to say such things as: '*x* endeavors to cause *y* to believe that *he*, *x*, believes with respect to *z* that it is *F*.'

Meaning to convey, then, presupposes the following concepts: making an utterance; causal contribution; perception; endeavor; thinking of; and belief. I will now say what it is to *address an utterance* for the *purpose* of conveying something. For simplicity, I will restrict myself to the situation wherein the speaker is addressing only one person; and I will use abbreviations introduced by previous definitions.

- D8 x addresses an utterance to z to convey the thought to z that y is $F =_{\text{Df}}$
- (1) x makes an utterance so that z will perceive that utterance and thereby believe that x thinks of y as F ;
 - (2) x does this to cause z to believe that he, x , intends to cause z to think of y as F ; and
 - (3) his purpose in doing this is to cause z to think of y as F .

When the conditions of the above definition are fulfilled, we may say that y is the *object* concerning which x means to convey something, and that the property of being F is the *content* of what it is that x means to convey with respect to y .

A broader concept may be obtained by revising the first clause of the definiens. Instead of saying that x intends to cause z to ‘believe that x thinks of y as F ’, we could say that x intends to cause z to ‘believe that there is a y such that he, x , thinks of y as being F ’. This would accommodate the situation (for example, an hallucination) wherein the belief has no intentional *object*.

SENSE AND REFERENCE

If a person x believes, with respect to a thing y , that y has a certain property P , then y is the *object* of the belief in question and the property P is the *content*. In terms of this intentional distinction between content and object, we may now spell out the linguistic distinction between sense and reference.⁵ The analysis of *sense* makes use of the concept of the *content* of thought, and the analysis of *reference*, or *designation*, makes use of the concept of the *object* of thought.

I will use the expression ‘attributive sense’ in order to distinguish the concept in question from the kind of sense that is sometimes attributed to proper names and demonstratives.

We begin with the ‘speaker’s sense’ of a predicate:

- D9 x uses P with the attributive sense $S =_{\text{Df}}$ x addresses an utterance for the purpose of conveying something, and P is that part of x ’s utterance which is intended to bring it about that S is the content of the thought he thus endeavors to cause.

‘Part’ is here to be read as ‘proper part’; hence if P is ‘that part of the utterance’ which is intended to bring about so-and-so, then there will be *another* part of the utterance, discrete from P , which is *not* intended to bring

about so-and-so. Hence P cannot be identified with the entire utterance.

The concept of the 'hearer's sense' is this:

- D10 z interprets x 's use of P as having the attributive sense $S =_{\text{Df}} z$ perceives that part of x 's utterance which is P and believes of it that x meant to use it with the attributive sense S .

When you said to me, 'that green chevrolet over there is out of control,' I interpreted the words 'out of control' (P) to express some such property S as this: being a car that is moving and not under anyone's control. And I assumed that you were attributing S to the thing you were trying to call my attention to.

I have spoken of the sense of predicative expressions – those utterances that are used to convey the *content* of one's thought. May we also ascribe a sense to such designative expressions as demonstratives and proper names?

We *could* ascribe a sense to proper names and demonstratives. For example, if I use a proper name in speaking to another person, then the demonstrative sense of that name on that occasion could be said to pertain to the relation or relations by means of which I then single out the object – or objects – of the belief I am expressing to the other person. Any such identifying relation will be a relation such that the bearer of the name is the thing to which the user of the name bears that relation. Consider again my statement, 'That green car is out of control.' One could say that the demonstrative sense of my expression 'that green car' might be the property expressed by 'the green car I have just been watching.' In this case, it would be the relational property of being someone who has just been watching one and only one green automobile.

The demonstrative sense of a name, then, would *not* be a property of the bearer of the name. It would be, rather, a relational property that the user of the name attributes to himself. The property, therefore, would be a property of the *user* of the name – provided there is a bearer of the name. And the corresponding relation would be one which the user of the name bears only to the bearer of the name.

But it is not clear that anything is to be gained by the introduction of this concept of demonstrative sense. For we do not need to use it in explicating the designative function of proper names and demonstratives.

What, then, of the concept of *designation*?

We first consider what may be called the 'speaker's designation' of a word:

- D11 x uses N to designate y =_{Df} x makes an utterance for the purpose of thereby conveying something about y ; and N is that part of x 's utterance which is intended to bring it about that y is the object of the thought that x thus endeavors to cause.

We have distinguished the case where one's thought *has* an object from the case where one's thought only *purports* to have an object. The latter situation may arise when I believe that there is one and only one person who is a devil and that that person is following me. Now it might be that I use the name 'Satan' for this person. But we cannot say that I use 'Satan' to *designate* him, for the person doesn't exist. Let us say that, in such a case, I 'mean' to use 'Satan' to designate something.' The relevant concept is this:

- D12 x means to use N to designate something =_{Df} x makes an utterance for the purpose of thereby conveying something about a certain thing; and N is that part of x 's utterance which is intended to bring it about that the thing, with respect to which he endeavors to convey this something, is the object of the thought he thus endeavors to convey.

The concept defined in D11 implies that defined in D12, but not conversely.

We now turn to what might be called the 'hearer's designation' of a word:

- D13 z interprets x 's use of N as designating y =_{Df} z perceives that part of x 's utterance which is N and believes of it that x meant to use it to designate y .

Or z may simply interpret x 's use of N as *purporting* to designate something. In such a case, z does not believe with respect to a thing y that x meant to use N to designate y ; z believes only that x meant to use N to designate something.

Looking back to our account of designation, one may ask 'How does it happen that x 's utterance of N can bring about the desired effect?' If x 's utterance is successful, then there will be someone z who is caused to perceive N and in consequence to think of y . There is, therefore, a *causal* factor that is involved: x 's utterance of N causes z to think of y . Hence there *can* be a 'causal theory of meaning': that is to say, there can be a causal *explanation* of the fact that a person's perception of an utterance n , or part of such an utterance, causes him to think of a certain thing y . But this is not a causal theory of *what* it is for one to think of another thing – much less a causal theory of what it is for one thing to designate another thing.

I have, then, attempted to single out certain intentional concepts in terms of which we may define *sense* and *reference*. We must distinguish (1) the sense in which we can say of a certain word in English that it has a fixed sense and that, at any time, it can be used in English only to designate one specific type of object at that time; and (2) the sense in which we can say of a certain word, say 'John', that it is now used to designate John and now used to designate Mary. We have been concerned here with (2), with 'speaker's meaning', and not with (1), not with 'linguistic meaning'. Linguistic meaning, so conceived, should be thought of as idealized speaker's meaning.

In this way, then, I would defend the principle of the primacy of the intentional.

NOTES

¹ I have defended a version of this view in *The First Person: An Essay on Reference and Intentionality* (Brighton and Minneapolis: The Harvester Press and The University of Minnesota Press, 1981).

² This example is adapted from one suggested by Keith Donnellan in 'Reference and Definite Descriptions', *Philosophical Review*, LXXV (1966) 281–304.

³ See Hector-Neri Castañeda, 'He: A Study in the Logic of Self-Consciousness', *Ratio*, 8 (1966) 130–157; 'Indicators and Quasi-Indicators', *American Philosophical Quarterly*, 4 (1967) 85–100.

⁴ Compare A.N. Whitehead's example of 'That college building is commodious', in *The Concept of Nature* (Cambridge: The University Press, 1930), pp. 6–7.

⁵ *Lecture on Ethics* (New York: Harper & Row, 1963), p. 226. The German reads: "...dann habe ich ihn nicht belogen, denn ich habe nicht deklariert, meine Gesinnung zu äußern." see Paul Menzer, ed., *Eine Vorlesung Kants über Ethik* (Berlin: Rolf Heise, 1925), p. 286. Compare Roderick M. Chisholm and Thomas D. Feehan, 'The Intent to Deceive', *Journal of Philosophy*, LXXIV (1977) 143–160.

Brown University
Providence, R.I.

INTENTIONALITY AND ITS PLACE IN NATURE

1.

Intentionality is that feature of certain mental states and events that consists in their (in a special sense of these words) being *directed at*, being *about*, being *of*, or *representing* certain other entities and states of affairs. If, for example, Robert has the belief that Ronald Reagan is President, then his belief is an intentional state because in the appropriate sense his belief is directed at, or about, or of, or represents Ronald Reagan and the state of affairs that Ronald Reagan is President. In such a case Ronald Reagan is the *intentional object* of Robert's belief, and the existence of the state of affairs that Ronald Reagan is President is the *condition of satisfaction* of his belief. If there is not anything that a belief is about, then it does not have an intentional object; and if the state of affairs it represents does not obtain, it is not satisfied.

Ascriptions of intentionality are of differing kinds, and as these differences have been a source of confusion, I will begin by sorting out some of them. Consider the statements made in utterances of the following sentences:

- A. Robert believes that Ronald Reagan is President.
- B. Bill sees that it is snowing.
- C. "Es regnet" means it's raining.
- D. My car thermostat perceives changes in the engine temperature.

Each of these statements ascribes intentionality, but the status of the ascriptions is different. A simply ascribes an intentional mental state, a belief, to a person; B does more than that, since to say of someone that he sees that something is the case implies not only that he has a certain form of intentionality but also that the intentionality is satisfied, i.e., that the conditions of satisfaction actually obtain. "See", like "know" but unlike "believe", is a success verb: x sees that p entails p . There is an intentional phenomenon reported by B, since in order to see something

one has to have a visual experience, and the visual experience is the vehicle of intentionality. But B does more than report a visual experience; it also reports that it is satisfied. Furthermore, visual experiences differ from beliefs in being conscious mental events rather than states. A man asleep can be literally said to believe that such and such, but not to see that such and such. Both beliefs and visual experiences are *intrinsic* intentional phenomena in the minds/brains of agents. To say that they are intrinsic is just to say that the states and events really exist in the minds/brains of agents; the ascription of these states and events is to be taken literally, not just as a manner of speaking, nor as shorthand for a statement describing some more complex set of events and relations going on outside the agents.

In this last respect the ascription of intentionality in A and B differs from C and D. C literally ascribes intentionality, though the intentionality is *not intrinsic*, but *derived*. It is literally true that the sentence “Es regnet” means it’s raining, but the intentionality in question is not intrinsic to the sentence. That very sentence might have meant something else or nothing at all. To ascribe this form of intentionality to it is shorthand for some statement or statements to the effect that speakers of German use the sentence literally to mean one thing rather than another, and the intentionality of the sentence is derived from this more basic form of intrinsic intentionality of speakers of German. In D, on the other hand, there is no literal ascription of intentionality at all because my car thermostat does not literally have any perceptions. D, unlike C, is a metaphorical ascription of intentionality; but, like C, its point depends on some intrinsic intentionality of agents. We use car thermostats to regulate engine temperatures and therefore they must be able to respond to changes in temperature. Hence the metaphor; and hence its harmlessness, provided we don’t confuse the analysis of A, B, and C, with that of D.

To summarize: even from this short list of statements there emerge several distinctions, which – in addition to the usual distinction between conscious and unconscious forms of intentionality – we will need to keep in mind.

1. The distinction between ascriptions of intentionality that imply that the intentional phenomenon is satisfied and those that do not, as illustrated by A and B.
2. The distinction between intentional states and intentional

- events, as also illustrated by A and B. (For brevity, I will sometimes call both "intentional states".)
3. The distinction between intrinsic intentionality and derived intentionality, as illustrated by the distinction between the intentionality ascribed in A and B, on the one hand, and C on the other.
 4. The distinction between literal ascriptions of intentionality such as A, B, and C, whose truth depends on the existence of some intentional phenomenon, whether intrinsic or derived; and metaphorical ascriptions, such as D, which do not literally ascribe any intentionality at all, even though the point of the metaphorical ascription may depend on some intrinsic intentionality of human agents.

In the rest of this paper I will deal only with intrinsic intentionality and the question I will discuss, to put it broadly is as follows: What is the place of intrinsic intentionality in nature?

2.

Intentional mental phenomena are part of our natural biological life history. Feeling thirsty, having visual experiences, having desires, fears, and expectations, are all as much a part of a person's biological life history as breathing, digesting, and sleeping. Intentional phenomena, like other biological phenomena, are real intrinsic features of certain biological organisms, in the same way that mitosis, meiosis, and the secretion of bile are real intrinsic features of certain biological organisms.

Intrinsic intentional phenomena are caused by neurophysiological processes going on in the brain, and they occur in and are realized in the structure of the brain. We do not know much about the details of how such things as neuron firings at synapses cause visual experiences and sensations of thirst; but we are not totally ignorant, and in the cases of these two intentional phenomena we even have pretty good evidence about their locations in the brain. That is, for at least some intentional phenomena, we have some idea of the special role of certain brain organs, such as the visual cortex or the hypothalamus, in producing the intentional phenomena. More important for our present discussion, our ignorance of how it all works in the brain is an empirical ignorance of

details and not the result of a metaphysical gulf between two incommensurable categories, the "Mind" and the "Body", which would prevent us from ever overcoming our ignorance. Indeed, the general sorts of relations involved between mental phenomena and the brain are quite familiar to us from other parts of nature. It is common in nature to discover that higher level features of a system are caused by the behavior of lower level microentities and realized in the structure of the system of microentities. For example, the solidity of the metal in the typewriter I am currently hammering on is caused by the behavior of the microparticles that compose the metal, and the solidity is realized in the structure of the system of microparticles, the atoms and molecules. The solidity is a feature of the system but not of any individual particle. Analogously, from what we know about the brain, mental states are features of the brain that are caused by the behavior of the elements at the microlevel and realized in the structure of the system of microelements, the neurons. A mental state is a feature of the system of neurons but not of any particular neuron. Furthermore, on this account there is no more reason for counting mental states as epiphenomenal than there would be for counting any other intrinsic, higher level features of the world, such as the solidity of this typewriter, as epiphenomenal.

In sum, certain organisms have intrinsic intentional states, these are caused by processes in the nervous systems of these organisms, and they are realized in the structure of these nervous systems. These claims should be understood in as naturalistic a sense as the claims that certain biological organisms digest food, that digestion is caused by processes in the digestive tract, and that it all goes on in the stomach and the rest of the digestive tract. Part of our difficulty in hearing the former claims naturalistically derives from the fact that the traditional vocabulary for discussing these problems is designed around a seventeenth century conception of the "mind/body problem". If we insisted on using the traditional jargon we might say: monism is perfectly compatible with dualism, provided it is a property dualism; and property dualism is compatible with complete physicalism, provided that we recognize that mental properties are just one kind of higher level property along with any number of other kinds. The view is not so much dualism as polyism, and it has the consequence that intrinsic mental properties are just one kind of higher level physical property among many others (which is perhaps a good reason for not using the traditional jargon at all).

It is a remarkable fact about contemporary intellectual life that the

existence of intrinsic intentional phenomena is frequently denied. It is sometimes said that the mind with its intentional states is something abstract, such as a computer program or a flow chart; or that mental states have no intrinsic *mental* status because they can be entirely defined in terms of their causes and effects; or that there aren't any *intrinsic* mental states, but rather that talk about mental states is just a manner of speaking that enables us to cope with our environment; and it is even sometimes said that mental terms should not be thought of as standing for actual things in the world at all. To catalogue the reasons that people have had for holding these views and denying everything that biology has to tell us about the brain would be to catalogue some of the leading intellectual confusions of the epoch. One, though only one, of the sources of confusion is the deeply held belief that if we grant the existence of intrinsic intentional states we will be confronted with an insoluble "mind/body" or "mind/brain" problem. But, to paraphrase Darwin, it is no more mysterious that the brain should cause mental phenomena than that bodies should have gravity. One might, and indeed one should, have a sense of awe and mystery in the face of both facts, but that sense would no more justify us in the denial of the existence of mental states than it would justify us in the denial of the existence of gravity.

Some philosophers feel that I am unjustified in simply asserting the existence of intrinsic intentional mental states and events in the world. For, they argue, might not the progress of science show them to be an illusion in the way that the appearance of the sun rising and setting over a stationary earth was shown to be an illusion? Isn't it just as prescientific to believe in intrinsic mental states as it is to believe that the earth is flat and in a fixed position in the universe?¹

But if we confine our attention for the moment to conscious intentional mental events and states – and they are, after all, the primary forms of intentionality – we can see that the analogy between the belief in a flat and fixed earth and the belief in the existence of mental phenomena breaks down. In the case of the earth there is a clear distinction between how things are and how they seem to be, but in the case of the very existence of conscious mental phenomena it is hard to know what a parallel distinction would look like. I know more or less exactly what it means to say that, though the earth seems flat and fixed, it is in fact not flat and fixed but rather is round and mobile; but I haven't the faintest idea what it would mean to say that, though it seems

to me that I am now conscious, in fact I am not really conscious but rather I am . . . What?

The reason we are unable to fill in the gap with anything that does not seem preposterous has been familiar since Descartes: If it consciously seems to me that I am conscious then that is enough for me to be conscious. And that is why there cannot be a general “how things seem”/“how they really are” distinction for the very existence of conscious mental states.

This is not, of course, to say that we cannot discover all sorts of surprising and counterintuitive things about our mental life, about the nature and mechanisms of both conscious and unconscious mental states. But it is to say that the distinction between how things seem and how they really are cannot apply to the *existence* of our own conscious mental phenomena.

3.

Since the resistance to treating consciousness and intentionality naturalistically, as just higher level properties among others, is so pervasive, and since the view of the place of intentionality in nature advanced in this article is so much out of step with what is currently accepted, I want to probe these issues a little more deeply. If one reads the standard literature on the “mind/body problem” over the past thirty years,² since the publication of Ryle’s *The Concept of Mind* (1949), one discovers a curious feature of this continuing dispute. Almost all the participants on both sides tacitly assume that the specifically mental features of conscious mental events cannot be ordinary physical features of the world like any other higher level features of the world. This assumption is often masked from us by the way theses are stated. Thus, when the identity theorist tells us that mental states just *are* brain states, there is a way of hearing that thesis which is perfectly consistent with our common-sense assumption of the intrinsic and irreducible character of consciousness and other forms of intentionality. We can hear the thesis as saying that mental processes are just processes going on in the brain in the way that digestive processes are processes going on in the digestive tract. But in general that is not what identity theorists are claiming. Under close scrutiny of the texts, particularly those parts of the texts where they are replying to dualist adversaries, it turns out that in general identity theorists (materialists, physicalists, functionalists,

etc.) end up by denying the existence of intrinsically mental features of the world. J. J. C. Smart, with typical candor, states the position clearly in responding to J. T. Stevenson:

My reply is that I do not admit that there are any such *P*-properties [i.e. properties of sensations that would prevent us from defining 'sensation' in terms of properties in a physicalist scheme]. (1970, p. 93)

Now why does Smart feel it necessary to deny what seems to Stevenson (and to me) an obvious truth? I believe it can only be because, in common with the tradition since Descartes, he thinks that to grant the reality of conscious mental phenomena is to grant the existence of some mysterious phenomena, some sort of "nomological danglers" beyond the reach of the physical sciences. On the other hand, consider those who challenge the tradition of materialism by asserting such obvious facts as that they are currently having a series of conscious states. They seem to think their claim commits them to some form of dualism, as if in asserting obvious facts about our waking life they are committed to the existence of some ontological category different from that of the ordinary physical world we all live in. One group of philosophers sees itself as defending the progress of science against residual superstitions. The other group sees itself as asserting obvious facts that any moment of introspection will reveal. But both accept the assumption that naive mentalism and naive physicalism must be inconsistent. Both accept the assumption that a purely physical description of the world could not mention any mental entities.

These are false assumptions. Unless one defines "physical" and "mental" in such a way that one is the negation of the other, there is nothing in our ordinary notions of mental phenomena and physical reality that excludes cases of the former from being instances of the latter.

Why then, to continue the investigation a step further, do both sides make this apparently obvious mistake? I think the answer must be that they take very seriously a whole tradition, going back at least to Descartes, with its endless disputes about substance, dualism, interaction, emergence, ontological categories, the freedom of the will, the immortality of the soul, the presuppositions of morality, and the rest of it. And in large measure this tradition revolves around the assumption that "mental" and "physical" name mutually exclusive categories. But suppose for a moment that we could forget all about this entire

tradition. Try to imagine that we are simply investigating the place in nature of our own human and animal mental states, intentional and otherwise, given what we know about biology, chemistry, and physics and what we know from our own experiences about our own mental states. I believe if we could forget the tradition, then the question as to the place of such states in nature would have an obvious answer. They are physical states of certain biochemical systems, viz., brains. But there is nothing reductive or eliminative about this view. Mental states with all their glorious or tiresome features – consciousness, intentionality, subjectivity, joy, anguish and the rest – are exactly as we knew they were all along.

Lest my view be misunderstood, I should like to state it with maximum simplicity. Take the most naive form of mentalism: There really are intrinsic mental states, some conscious, some unconscious; some intentional, some nonintentional. As far as the conscious ones are concerned they pretty much have the mental properties they seem to have, because in general for such properties there is no distinction between how things are and how they seem. Now take the most naive version of physicalism: The world consists entirely of physical particles, including the various sorts of relations between them. As far as real things in the world are concerned there are only physical particles and various arrangements of physical particles. Now, my point is that it is possible to accept both of these views exactly as they stand, without any modification whatever. Indeed the first is simply a special case of the second.

4.

Granted that intentional mental states really do exist and are not to be explained away as some kind of illusion or eliminated by some sort of redefinition, what role do they play in a naturalistic or scientific description of nature?

Just as it is a biological fact that certain sorts of organisms have certain sorts of mental states, so it is equally a biological fact that certain mental states function causally in the interactions between the organism and the rest of nature and in the production of the behavior of the organism. It is just a fact of biology that sometimes thirst will cause an organism to drink water, that hunger will cause it to seek and consume food, and that sexual desire will cause it to copulate. In the

case of human beings, at a much more sophisticated though equally biological level, the beliefs a person has about what is in his or her economic interest may play a causal role in how he or she votes in political elections, literary preferences may play a causal role in the purchase and reading of books, and the desire to be someplace other than where one is may play a causal role in a person's buying a plane ticket, driving a car, or getting on a bus. Though the fact of causal relations involving intentional mental states is pretty obvious, what is a great deal less obvious is the logical structure of the causal relations involved and the consequent implications that those causal relations have for the logical structure of the explanation of human behavior.

Explanations involving intentionality have certain logical features not common to explanations in the other physical sciences. The first of these is the specific role of intentional causation in the production of certain sorts of animal and human behavior. The essential feature of intentional causation is that the intentional state itself functions causally in the production of its own conditions of satisfaction or its conditions of satisfaction function causally in its production. In the one case the representation, as a representation, produces what it represents; in the other case the object or state of affairs represented functions causally in the production of its representation. This point can be made clear by considering some examples. If I now have a strong desire to drink a cup of coffee and I act on that desire so as to satisfy it, then the desire whose content is

(that I drink a cup of coffee)

causes the very state of affairs, that I drink a cup of coffee. Now, in this simple and paradigmatic case of intentional causation, the desire represents the very state of affairs that it causes. The much discussed "internal connection" between "reasons for action" and the actions that they cause is just a reflection of this underlying feature of intentional causation. Since the cause is a representation of that which it causes, the specification of the cause, as cause, is indirectly already a specification of the effect.

Sometimes, indeed, the intentional state has as part of its conditions of satisfaction, as part of its intentional content, that it must function causally if it is to be satisfied. Thus, for example, intentions can only be satisfied if the actions that they represent are caused by the intentions that represent them. In this respect intentions differ from desires: a

desire can be satisfied even if it does not cause the conditions of its satisfaction; whereas an intention can be satisfied only if it causes the rest of its own conditions of satisfaction. For example, if I want to be rich and I become rich, my desire will be satisfied even if that desire played no causal role in my becoming rich; but if I intend to earn a million dollars and I wind up with a million dollars quite by accident, in such a way that my intention to earn the money played no causal role consciously or unconsciously in my getting it, then although the state of affairs represented by my intention came about, the intention itself was not satisfied, i.e., the intention was never carried out. Intentions, unlike desires, have intentional causation built into their intentional structure; they are causally self-referential in the sense that they can only be satisfied *if they cause* the very action they represent. Thus, the prior intention to drink a cup of coffee differs in its content from the desire to drink a cup of coffee, as we can see by contrasting the following representation of the conditions of satisfaction of a prior intention with our representation above of the conditions of satisfaction of the corresponding desire:

(that I drink a cup of coffee and that this prior intention cause that I drink a cup of coffee).

Cases of “volition”, such as desires and intentions, have what I call the “world-to-mind direction of fit” (the aim of the state is to get the world to change to match the content of the desire or intention) and the “mind-to-world direction of causation” (the mental state causes the state of affairs in the world that it represents). Cases of “cognition”, such as perception, memory, and belief, function conversely as far as direction of fit and intentional causation are concerned. Thus, they have the mind-to-world direction of fit (the aim of the mental state is not to create a change in the world, but to match some independently existing reality); and, where intentional causation functions in the production of the intentional state, they have the world-to-mind direction of causation (if I correctly perceive or remember how things are in the world, then their being that way must cause my perceiving them or remembering them as being that way).

5.

I believe that a full account of the role of intentionality and its place in

nature requires much more study of intentional causation than has yet been done or than I can undertake in this essay. But by way of giving the reader some idea of the importance of intentional causation, I want to mention just three of the implications of this brief sketch of intentional causation for a causal account of human and animal behavior and for ways in which such a causal account differs from certain standard models of what we have as canonical explanations in the usual natural sciences.

1. In any causal explanation, the propositional content of the explanation specifies a cause. But in intentional explanations the cause specified is itself an intentional state with its own propositional content. The canonical specification, therefore, of the cause in an intentional explanation doesn't just *specify* the propositional content of the cause, but it must actually *repeat* in the explanation (at least some of) the propositional content that is functioning causally in the operation of the cause. So, for example, if I buy a plane ticket because I want to go to Rome, then in the explanation:

I did it because I want to go to Rome.

I repeat the very propositional content functioning in the operation of the desire:

I want to go to Rome.

Intentional explanations are more or less adequate as they accurately repeat in the explanation the propositional content functioning in the cause itself. It is a further consequence of this feature that the concepts used in the canonical form of the explanation don't just describe a cause; rather, the very concepts themselves must function in the operation of the cause. So, if I say that a man voted for Reagan because he thought it would increase the probability that he would be rich and happy, such concepts as being rich and being happy can be used in the explanation to specify a cause only if they also function as part of the cause.

These features have no analogue in the standard physical sciences. If I explain the rate of acceleration of a falling body in terms of gravitational attraction together with such other forces as friction operating on the body, the propositional content of my explanation makes reference to features of the event such as gravity and friction, but the features themselves are not propositional contents or parts of

propositional contents.

This is a familiar point in the history of discussions of the nature of psychological explanation, but it seems to me that it has not been properly stated or appreciated. I believe it is part of what Dilthey (1962) was driving at when he said that the method of *Verstehen* was essential to the social sciences, and it was part of what Winch (1958) was driving at when he said that concepts used in the explanation of human behavior must also be concepts that are available to the agent whose behavior is being explained. I think an analysis of intentional causation would provide us with a deeper theoretical understanding of the points that Dilthey and Winch were after.

2. Statements of intentional causation do not require the statement of a covering law in order to be validated or in order to be causally explanatory. In a subject like physics we assume that no causal explanation of a phenomenon is fully explanatory unless it can be shown to instantiate some general law or laws. But in the case of intentional causation this is not generally the case. Even if we believe that there are laws, stateable in some terms or other, which any given instance of behavior instantiates, it is not essential to giving a causal explanation of human behavior in terms of intentional causation that we be able to state any such laws or even believe that there are such laws.

3. Teleological forms of explanation are those in which a phenomenon is explained in terms of goals, aims, purposes, intentions, and similar phenomena. If teleological explanation is really a subclass of scientific explanation, it would appear that nature must actually contain teleological phenomena. The account of intentionality and its place in nature that I have been urging has the consequences both that nature contains teleological phenomena and that teleological explanations are the appropriate forms of explanation for certain sorts of events. Indeed, it is an immediate logical consequence of the claim that goals, aims, purposes, and intentions are intrinsic features of certain biological organisms that teleology is an intrinsic part of nature, for by definition such phenomena are teleological. And it is an immediate consequence of the characterization I have given of these phenomena that teleological explanations are the appropriate forms for explaining certain sorts of events, since these phenomena cause events by way of the form of intentional causation that is peculiar to teleology.

All the states I have called "teleological" have the world-to-mind direction of fit and the mind-to-world direction of causation. The

explanatory role of citing such states in teleological explanations can best be illustrated by example. Consider the case of an animal, say a lion, moving in an erratic path through tall grass. The behavior of the lion is explicable by saying that it is stalking a wildebeest, its prey. The stalking behavior is caused by a set of intentional states: it is *hungry*, it *wants* to eat the wildebeest, it *intends* to follow the wildebeest with the *aim* of catching, killing, and eating it. Its intentional states represent possible future states of affairs; they are satisfied only if those states of affairs come to pass (world-to-mind direction of fit); and its behavior is an attempt to bring about those states of affairs (mind-to-world direction of causation). The claim that teleology is part of nature amounts to the claim that certain organisms contain future-directed intentional states with the world-to-mind direction of fit, and that these states are capable of functioning causally to bring about their conditions of satisfaction.

It is worth emphasizing the logical features of teleological explanation because on some accounts a teleological explanation explains an event by the occurrence of a future event, as if, for example, the eating of the prey explained the stalking behavior.³ But on my account this conception has things back to front. All valid teleological explanations are species of explanation in terms of intentional causation, and there is no mysterious backwards operation of intentional causation. The stalking behavior at time t_1 is explained by present and prior intentional states at t_1 and t_0 , and these aim at the eating behavior of t_2 .

In the great scientific revolution of the seventeenth century the rejection of teleology in physics was a liberating step. Again in the great Darwinian revolution of the nineteenth century the rejection of a teleological account of the origins of the species was a liberating step. In the twentieth century there has been an overwhelming temptation to complete the picture by rejecting teleology in the sciences of man. But ironically the liberating move of the past has become constraining and counterproductive in the present. Why? Because it is just a plain fact about human beings that they do have desires, goals, intentions, purposes, aims, and plans, and these play a causal role in the production of their behavior. Those human sciences in which these facts are simply taken for granted, such as economics, have made much greater progress than those branches, such as behavioristic psychology, which have been based on an attempted denial of these facts. Just as it was bad science to treat systems that lack intentionality as if they had it, so it is

equally bad science to treat systems that have intrinsic intentionality as if they lacked it.

NOTES

¹ Rorty (forthcoming), p. 84.

² I am thinking of the sort of articles to be found in Borst (1970), Rosenthal (1971), and Block (1980).

³ For a discussion of this conception see Braithwaite (1953), Chapter X.

REFERENCES

- Block, N. (ed.): 1980, *Readings in Philosophical Psychology*, Vol. 1, Harvard University Press, Cambridge, Mass.
- Borst, C. V. (ed.): 1970, *The Mind-Brain Identity Theory*, Macmillan, London.
- Braithwaite, R. B.: 1953, *Scientific Explanation*, Cambridge University Press, Cambridge, England.
- Dilthey, W.: 1962, *Meaning in History*, H. P. Rickman, (trans. and ed.), New York.
- Rorty, R.: forthcoming, 'Mind as Ineffable', in R. Q. Elvee (ed.), *Mind and Nature*, Harper & Row, San Francisco.
- Rosenthal, D. M. (ed.): 1971, *Materialism and the Mind-Body Problem*, Prentice-Hall, Englewood Cliffs, N.J.
- Ryle, G.: 1949, *The Concept of Mind*, Hutchinson's University Library, London.
- Smart, J. J. C.: 1970, 'Further Remarks on Sensations and Brain Processes', in Borst (1970), pp. 93-94.
- Winch, P.: 1958, *The Idea of a Social Science*, Routledge & Kegan Paul, London.

Department of Philosophy
University of California, Berkeley
Berkeley, California 94720
U.S.A.

PART V

EPISTEMOLOGY AND COGNITION

WHY REASON CAN'T BE NATURALIZED*

The preceding lecture described the failure of contemporary attempts to "naturalize" metaphysics; in the present lecture I shall examine attempts to naturalize the fundamental notions of the theory of knowledge, for example the notion of a belief's being *justified* or *rationally acceptable*.

While the two sorts of attempts are alike in that they both seek to reduce "intentional" or mentalistic notions to materialistic ones, and thus are both manifestations of what Peter Strawson has recently described as a permanent tension in philosophy,¹ in other ways they are quite different. The materialist metaphysician often uses such traditional metaphysical notions as *causal power*, and *nature* quite uncritically (I have also read papers in which one finds the locution "realist truth", as if everyone understood this notion except a few fuzzy anti-realists). The "physicalist" generally doesn't seek to *clarify* these traditional metaphysical notions, but just to show that science is progressively verifying the *true* metaphysics. That is why it seems just to describe his enterprise as "natural metaphysics", in strict analogy to the "natural theology" of the 18th and 19th centuries. Those who raise the slogan "epistemology naturalized", on the other hand, generally *disparage* the traditional enterprises of epistemology. In this respect, moreover, they do not differ from philosophers of a less reductionist kind; the criticism they voice of traditional epistemology – that it was in the grip of a "quest for certainty", that it was unrealistic in seeking a "foundation" for knowledge as a whole, that the "foundation" it claimed to provide was by no means indubitable in the way it claimed, that the whole "Cartesian enterprise" was a mistake, etc., etc., – are precisely the criticisms one hears from philosophers of all countries and types. Hegel already denounced the idea of an "Archimedean point" from which epistemology could judge all of our scientific, legal, moral, religious, etc. beliefs (and set up standards for all of the special subjects). It is true Russell and Moore ignored these strictures of Hegel (as they ignored Kant), and revived "foundationalist epistemology"; but today that enterprise has few

defenders. The fact that the naturalized epistemologist is trying to reconstruct what he can of an enterprise that few philosophers of any persuasion regard as unflawed is perhaps the explanation of the fact that the naturalistic tendency in epistemology expresses itself in so many incompatible and mutually divergent ways, while the naturalistic tendency in metaphysics appears to be and regards itself as a unified movement.

EVOLUTIONARY EPISTEMOLOGY

The simplest approach to the problem of giving a naturalistic account of reason is to appeal to Darwinian evolution. In its crudest form, the story is familiar: reason is a capacity we have for discovering truths. Such a capacity has survival value; it evolved in just the way that any of our physical organs or capacities evolved. A belief is rational if it is arrived at by the exercise of this capacity.

This approach assumes, at bottom, a metaphysically “realist” notion of truth – truth as “correspondence to the facts” or something of that kind. And this notion, I have argued,² is incoherent. We don’t have notions of the “existence” of things or of the “truth” of statements that are independent of the versions we construct and of the procedures and practices that give sense to talk of “existence” and “truth” within those versions. Do *fields* “exist” as physically real things? Yes, fields really exist – relative to one scheme for describing and explaining physical phenomena; relative to another there are particles, plus “virtual” particles, plus “ghost” particles, plus . . . Is it true that *brown* objects exist? Yes, relative to a common sense version of the world – although one cannot give a necessary and sufficient condition for an object³ to be brown, (one that applies to all objects, under all conditions) in the form of a finite closed formula in the language of physics. Do *dispositions* exist? Yes, in our ordinary way of talking (although disposition-talk is just as recalcitrant to translation into physicalistic language as counterfactual talk, and for similar reasons). We have many irreducibly different but legitimate ways of talking, and true “existence” statements in all of them.

To postulate a set of “ultimate” objects, the Furniture of the World, or what you will, whose “existence” is *absolute*, not relative to our discourse at all, and a notion of truth as “correspondence” to

these Ultimate Objects is simply to revive the whole failed enterprise of traditional metaphysics. *How unsuccessful attempts to revive that enterprise have been we saw in the last lecture.*

Truth, in the only sense in which we have a vital and working notion of it, is rational acceptability (or, rather, rational acceptability under sufficiently good epistemic conditions; and which conditions are epistemically better or worse is relative to the type of discourse in just the way rational acceptability itself is). But to substitute this characterization of truth into the formula “reason is a capacity for discovering truths” is to see the emptiness of that formula at once: “reason is a capacity for discovering what is (or would be) rationally acceptable” is *not* the most informative statement a philosopher might utter. The evolutionary epistemologist must either presuppose a “realist” (i.e., a metaphysical) notion of truth or see his formula collapse into vacuity.

Roderick Firth⁴ has argued that, in fact, it collapses into a kind of epistemic vacuity on *any* theory of rational acceptability (or truth). For, he points out, whatever we take the correct epistemology (or the correct theory of truth) to be, we have no way of identifying truths except to posit that the statements that are currently rationally acceptable (by our lights) are true. Even if these beliefs are false, even if our rational beliefs contribute to our survival for some reason *other* than truth, the way “truths” are identified *guarantees* that reason will seem to be a “capacity for discovering truths”. This characterization of reason has thus no real empirical content.

The evolutionary epistemologist could, I suppose, try using some notion *other* than the notion of “discovering truths”. For example, he might try saying that “reason is a capacity for arriving at beliefs which *promote our survival*” (or our “inclusive genetic fitness”). But this would be a loser! Science itself, and the methodology which we have developed since the 17th century for constructing and evaluating theories, has *mixed* effects on inclusive genetic fitness and all too uncertain effects on survival. If the human race perishes in a nuclear war, it may well be (although there will be no one alive to say it) that scientific beliefs did *not*, in a sufficiently long time scale, promote “survival”. Yet that will not have been because the scientific theories were not rationally acceptable, but because our *use* of them was irrational. In fact, if rationality were measured by survival-value, then the proto-beliefs of the cockroach, who has been around for tens of

millions of years longer than we, would have a far higher claim to rationality than the sum total of human knowledge. But such a measure would be cockeyed; there is no contradiction in imagining a world in which people have utterly irrational beliefs which for some reason enable them to survive, or a world in which the most rational beliefs quickly lead to extinction.

If the notion of "truth" in the characterization of rationality as a "capacity for discovering truths" is problematic, so, almost equally, is the notion of a "capacity". In one sense of the term, *learning* is a "capacity" (even, a "capacity for discovering truths"), and *all* our beliefs are the product of *that* capacity. Yet, for better or worse, not all our beliefs are rational.

The problem here is that there are no sharp lines in the brain between one "capacity" and another (Chomskians to the contrary). Even seeing includes not just the visual organs, the eyes, but the whole brain; and what is true of seeing is certainly true of *thinking* and *inferring*. We draw lines between one "capacity" and another (or build them into the various versions we construct); but a sharp line at one level does not usually correspond to a sharp line at a lower level. The table at which I write, for example, is a natural unit at the level of everyday talk; I am aware that the little particle of food sticking to its surface (I must do something about that!) is not a "part" of the table; but at the physicist's level, the decision to consider that bit of food to be outside the boundary of the table is not natural at all. Similarly, "believing" and "seeing" are quite different at the level of ordinary language psychology (and usefully so); but the corresponding brain-processes interpenetrate in complex ways which can only be separated by looking outside the brain, at the environment and at the output behavior *as structured by our interests and saliences*. "Reason is a capacity" is what Wittgenstein called a "grammatical remark"; by which he meant (I think) not an analytic truth, but simply the sort of remark that philosophers often *take* to be informative when in fact it tells us nothing useful.

None of this is intended to deny the obvious scientific facts: that we would not be able to reason if we did not have brains, and that those brains are the product of evolution by natural selection (I trust I am allowed to say that even in the state of California!) What is wrong with evolutionary epistemology is not that the scientific facts are wrong, but that they don't answer any of the philosophical questions.

THE RELIABILITY THEORY OF RATIONALITY

A more sophisticated recent approach to these matters, proposed by Professor Alvin Goldman,⁵ runs as follows: let us call a *method* (as opposed to a single belief) *reliable* if the method leads to a high frequency (say, 95%) of *true* beliefs in a long run series of representative applications (or *would* lead to such a high truth-frequency in such a series of applications). Then (the proposal goes) we can define a *rational* belief to be one which is *arrived at by using a reliable method*.

This proposal does not avoid the first objection we raised against evolutionary epistemology: it too presupposes a metaphysical notion of truth. Forgetting that rational acceptability does the lion's share of the work in fixing the notion of "truth", the reliability theorist only pretends to be giving an analysis of rationality in terms that do not presuppose it. The second objection we raised against evolutionary epistemology viz. that the notion of a "capacity" is hopelessly vague and general, is met, however, by replacing that notion with the notion of an arbitrary method for generating true or false statements, and then restricting the class to those methods (in this sense) whose reliability (as defined) is high. "Learning" may be a method for generating statements, but its *reliability* is not high enough for every statement we "learn" to count as rationally acceptable, on this theory. Finally, *no* hypothesis is made as to whether the reliable methods we employ are the result of biological evolution, cultural evolution, or what: this is regarded as no part of the theory of what rationality is, in this account.

This account is vulnerable to many counter examples, however. *One* is the following: suppose that Tibetan Buddhism is, in fact, *true*, and that the Dalai Lama is, in fact, *infallible* on matters of faith and morals. Anyone who believes in the Dalai Lama, and who invariably believes any statement the Dalai Lama makes on a matter of faith or morals, follows a method which is 100% reliable; thus, if the reliability theory of rationality were correct, such a person's beliefs on faith and morals would all be rational *even if his argument for his belief that the Dalai Lama is never wrong is "the Dalai Lama says so"*.

CULTURAL RELATIVISM

I have already said that, in my view, truth and rational acceptability – a claim's being right and someone's being in a position to make it –

are relative to the sort of language we are using and the sort of context we are in. "That weighs one pound" may be true in a butcher shop, but the same sentence would be understood very differently (as demanding four decimal places of precision, perhaps) if the same object were being weighed in a laboratory. This does not mean that a claim is right *whenever* those who employ the language in question would accept it as right in its context, however. There are two points that must be *balanced*, both points that have been made by philosophers of many different kinds: (1) talk of what is "right" and "wrong" in any area only makes sense against the background of an *inherited tradition*; but (2) traditions themselves can be *criticized*. As Austin says, remarking on a special case of this,⁶ "superstition and error and fantasy of all kinds do become incorporated in ordinary language and even sometimes stand up to the survival test (only, when they do, why should we not detect it?)".

What I am saying is that the "standards" accepted by a culture or a subculture, either explicitly or implicitly, cannot *define* what reason is, even in context, because they *presuppose* reason (reasonableness) for their interpretation. On the one hand, there is no notion of reasonableness at all *without* cultures, practices, procedures; on the other hand, the cultures, practices, procedures we inherit are not an algorithm to be slavishly followed. As Mill said, commenting on his own inductive logic, there is no rule book which will not lead to terrible results "if supposed to be conjoined with universal idiocy". Reason is, in this sense, both immanent (not to be found outside of concrete language games and institutions) and transcendent (a regulative idea that we use to criticize the conduct of *all* activities and institutions).

Philosophers who lose sight of the immanence of reason, of the fact that reason is always relative to context and institution, become lost in characteristic philosophical fantasies. "The ideal language", "inductive logic", "the empiricist criterion of significance" – these are the fantasies of the Positivist, who would replace the vast complexity of human reason with a kind of intellectual Walden II. The Absolute Idea – this is the fantasy of Hegel, who, without ignoring that complexity, would have us (or, rather, "spirit") reach an end-stage at which we (it) could comprehend it all. Philosophers who lose sight of the transcendence of reason become cultural (or historical) relativists.

I want to talk about cultural relativism, because it is one of the

most influential – perhaps the most influential – forms of naturalized epistemology extant, although not usually recognized as such.

The situation is complicated, because cultural relativists usually *deny* that they are cultural relativists. (I shall count a philosopher as a cultural relativist for our purposes if I have not been able to find anyone who can explain to me why the man *isn't* a cultural relativist.) Thus I count Richard Rorty as a cultural relativist, because his explicit formulations are relativist ones (he identifies truth with right assertibility by the standards of one's cultural peers, for example), and because his entire attack on traditional philosophy is mounted on the basis that the nature of reason and representation are non-problems, because the only kind of truth it makes sense to seek is to convince one's cultural peers. Yet he himself *tells* us that relativism is self-refuting.⁷ And I count Michel Foucault as a relativist because his insistence on the determination of beliefs by language is so overwhelming that it is an incoherence on his part not to apply his doctrine to his *own* language and thought. (Whether Heidegger ultimately escaped something very much like cultural, or rather historical, relativism is an interesting question.)

Cultural relativists are not, in their own eyes, scientific or “physicalistic”. They are likely to view materialism and scientism as just the hang-ups of one particular cultural epoch. If I count them as “naturalized epistemologists” it is because their doctrine is, nonetheless, a product of the same deference to the claims of Nature, the same desire for harmony with the world-version of some science, as physicalism. The difference in style and tone is thus explained: the physicalist's paradigm of science is a *hard* science, *physics* (as the term “physicalism” suggests); the cultural relativist's paradigm is a *soft* science – anthropology, or linguistics, or psychology, or history, as the case may be. That reason is whatever the norms of the local culture determine it to be is a naturalist view inspired by the *social* sciences, including history.

There is something which makes cultural relativism a far more dangerous cultural tendency than materialism. At bottom, there is a deep irrationalism to cultural relativism, a denial of the possibility of *thinking* (as opposed to making noises in counterpoint or in chorus). An aspect of this which is of special concern to philosophy is the suggestion, already mentioned, that the deep questions of philosophy are not deep at all. A corollary to this suggestion is that philosophy,

as traditionally understood, is a *silly* enterprise. But the questions *are* deep, and it is the easy answers that are silly. Even seeing that relativism is inconsistent is, if the knowledge is taken seriously, seeing something important about a deep question. Philosophers *are* beginning to talk about the great issues again, and to feel that something can be *said* about them, even if there are no grand or ultimate solutions. There is an excitement in the air. And if I react to Professor Rorty's book⁸ with a certain sharpness, it is because one more "deflationary" book, one more book telling us that the deep questions aren't deep and the whole enterprise was a mistake, is just what we *don't* need right now. Yet I am grateful to Rorty all the same, for his work has the merit of addressing profound questions head-on.

So, although we all know that cultural relativism is inconsistent (or say we do) I want to take the time to say again that it is inconsistent. I want to point out one reason that it is – not one of the quick, logic-chopping refutations (although every refutation of relativism teaches us something about reason) but a somewhat messy, somewhat "intuitive", reason.

I shall develop my argument in analogy with a well known argument against "methodological solipsism". The "methodological solipsist" – one thinks of Carnap's *Logische Aufbau* or Mach's *Analyse der Empfindungen* – holds that *all* our talk can be reduced to talk about experiences and logical constructions out of experiences. More precisely, he holds that everything he can conceive of is identical (in the ultimate logical analyses of his language) with one or another complex of his *own* experiences. What makes him a *methodological* solipsist as opposed to a real solipsist is that he kindly adds that *you*, dear reader, are the "I" of this construction when *you* perform it – he says *everybody* is a (methodological) solipsist.

The trouble, which should be obvious, is that his two stances are ludicrously incompatible. His solipsist stance implies an enormous assymetry between persons: my body is a construction out of my experiences, in the system, but *your* body isn't a construction out of *your* experiences. It's a construction out of *my* experiences. And *your* experiences – viewed from within the system – are a construction out of *your* bodily behavior, which, as just said, is a construction out of *my* experiences. My experiences are different from everyone else's (within the system) in that they are what

everything is constructed from. But his transcendental stance is that it's all symmetrical – the “you” he addresses his higher order remark to cannot be the *empirical* “you” of the system. But if it's really true that the “you” of the system is the only “you” he can *understand*, then the transcendental remark is *unintelligible*. Moral: don't be a methodological solipsist unless you are a *real* solipsist!

Consider now the position of the cultural relativist who says, “When I say something is *true*, I mean that it is correct according to the norms of *my* culture.” If he adds, “When a member of a different culture says that something is true, what he means (whether he knows it or not) is that it is in conformity with the norms of *his* culture,” then he is in exactly the same plight as the methodological solipsist.

To spell this out, suppose R. R., a cultural relativist, says

When Karl says ‘Schnee ist weiss’, what Karl means (whether he knows it or not) is that snow is white *as determined by the norms of Karl's culture*

(which we take to be German culture).

Now the sentence “Snow is white as determined by the norms of German culture” is itself one which R. R. has to *use*, not just mention, to say what Karl says. On his own account, what R. R. means by *this* sentence is

“Snow is white as determined by the norms of German culture” is true by the norms of R. R.'s culture

(which we take to be American culture).

Substituting this back into the first displayed utterance, (and changing to indirect quotation) yields:

When Karl says “Schnee ist weiss”, what he means (whether he knows it or not) is that it is true as determined by the norms of American culture that it is true as determined by the norms of German culture that snow is white.

In general, if R. R. understands *every* utterance *p* that *he* uses as meaning “it is true by the norms of American culture that *p*”, then he must understand his own hermeneutical utterances, the utterances he

uses to interpret others, the same way, no matter how many qualifiers of the “according to the norms of German culture” type or however many footnotes, glosses, commentaries on the cultural differences, or whatever, he accompanies them by. Other cultures become, so to speak, logical constructions out of the procedures and practices of American culture. If he now attempts to add “the situation is reversed from the point of view of the *other* culture” he lands in the predicament the methodological solipsist found himself in: the transcendental claim of a *symmetrical* situation cannot be *understood* if the relativist doctrine is right. And to say, as relativists often do, that the other culture has “incommensurable” concepts is no better. This is just the transcendental claim in a special jargon.

Stanley Cavell⁹ has recently written that scepticism about other minds can be a significant problem because we don’t, in fact, always fully acknowledge the reality of others, their equal *validity* so to speak. One might say that the methodological solipsist is led to his transcendental observation that everyone is equally the “I” of the construction by his praiseworthy desire to *acknowledge* others in this sense. But you *can’t* acknowledge others in this sense, which involves recognizing that the situation *really is* symmetrical, if you think they are really constructions out of *your* sense-data. Nor can you acknowledge others in this sense if you think that the *only* notion of truth there is for *you* to understand is “truth-as-determined-by-the-norms-of-*this*-culture”.

For simplicity, I have discussed relativism with respect to truth, but the same discussion applies to relativism about rational acceptability, justification, etc; indeed, a relativist is unlikely to be a relativist about one of these notions and not about the others.

CULTURAL IMPERIALISM

Just as the methodological solipsist can become a *real* solipsist, the cultural relativist can become a cultural imperialist. He can say, “Well then, truth – the only notion of truth I understand – is defined by the norms of *my* culture.” (“After all,” he can add, “which norms should I rely on? The norms of *somebody else’s* culture?”) Such a view is no longer relativist at all. It postulates an *objective* notion of truth – although one that is said to be a product of our culture, and to be defined by our culture’s criteria (I assume the cultural imperialist is

one of *us*). In this sense, just as consistent solipsism becomes indistinguishable from realism (as Wittgenstein said in the *Tractatus*), consistent cultural relativism also becomes indistinguishable from realism. But cultural imperialist realism is a special *kind* of realism.

It is realist in that it accepts an objective difference between what is true and what is merely thought to be true. (Whether it can consistently account for this difference is another question.)

It is not a *metaphysical* or transcendental realism, in that truth cannot go beyond right assertibility, as it does in metaphysical realism. But the notion of right assertibility is fixed by "criteria", in a positivistic sense: something is rightly assertible only if the norms of the culture specify that it is; these norms are, as it were, an *operational definition* of right assertibility, in this view.

I don't know if any philosopher holds such a view, although several philosophers have let themselves fall into talking at certain times as if they did. (A philosopher in this mood is likely to say, "X is *our* notion," with a certain petulance, where X may be *reason*, *truth*, *justification*, *evidence*, or what have you.)

This view is, however, self-refuting, at least in our culture. I have discussed this elsewhere;¹⁰ the argument turns on the fact that our culture, unlike totalitarian or theocratic cultures, does not have "norms" which decide *philosophical* questions. (Some philosophers have thought it does; but they had to postulate a "depth grammar" accessible only to *them*, and not describable by ordinary linguistic or anthropological investigation.) Thus the philosophical statement:

A statement is true (rightly assertible) only if it is assertible according to the norms of modern European and American culture

is itself neither assertible nor refutable in a way that requires assent by everyone who does not deviate from the norms of modern European and American culture. So, if this statement is true, it follows that it is not true (not rightly assertible). Hence it is not true Q.E.D. (I believe that *all* theories which identify truth or right assertibility with what people agree with, or with what they would agree with in the long run, or with what educated and intelligent people agree with, or with what educated and intelligent people would agree with in the long run, are contingently self-refuting in this same way.)

Cultural imperialism would not be contingently self-refuting in this way if, as a matter of contingent fact, our culture were a totalitarian culture which erected its own cultural imperialism into a required dogma, a culturally normative belief. But it would still be wrong. For every culture has norms which are vague, norms which are unreasonable, norms which dictate inconsistent beliefs. We have all become aware how many inconsistent beliefs about *women* were culturally normative until recently, and are still strongly operative, not only in subcultures, but in all of us to some extent; and examples of inconsistent but culturally normative beliefs could easily be multiplied. Our task is not to mechanically *apply* cultural norms, as if they were a computer program and we were the computer, but to interpret them, to criticize them, to bring them and the ideals which inform them into reflective equilibrium. Cavell has aptly described this as “confronting the culture with itself, along the lines in which it meets in me”. And he adds,¹¹ “This seems to me a task that warrants the name of Philosophy.” In this sense, we are all called to be philosophers, to a greater or lesser extent.

The culturalist, relativist or imperialist, like the historicist, has been caught up in the fascination of something really fascinating; but caught up in a sophomorish way. Traditions, cultures, history, deserve to be emphasized, as they are not by those who seek Archimedian points in metaphysics or epistemology. It is true that we speak a public language, that we inherit versions, that talk of truth and falsity only make sense against the background of an “inherited tradition”, as Wittgenstein says. But it is also true that we constantly remake our language, that we make new versions out of old ones, and that we have to use reason to do all this, and, for that matter, even to understand and apply the norms we do not alter or criticize. Consensus definitions of reason do not work, because consensus among grown-ups *presupposes* reason rather than defining it.

QUINIAN POSITIVISM

The slogan ‘Epistemology Naturalized’ is the title of a famous paper by Quine.¹² If I have not discussed that paper up to now, it is because Quine’s views are much more subtle and much more elaborate than

the disastrously simple views we have just reviewed, and it seemed desirable to get the simpler views out of the way first.

Quine's philosophy is a large continent, with mountain ranges, deserts, and even a few Okefenokee Swamps. I do not know how all of the pieces of it can be reconciled, if they can be; what I shall do is discuss two different strains that are to be discerned in Quine's epistemology. In the present section I discuss the positivistic strain; the next section will discuss 'Epistemology Naturalized'.

The positivist strain, which occurs early and late, turns on the notion of an *observation sentence*. In his earliest writings, Quine gave this a phenomenalist interpretation, but since the 1950's, at least, he has preferred a definition in neurological and cultural terms. First, a preliminary notion: The *stimulus meaning* of a sentence is defined to be the set of stimulations (of "surface neurons") that would "prompt assent" to the sentence. It is thus supposed to be a *neurological correlate* of the sentence. A sentence may be called "stimulus-true" for a speaker if the speaker is actually experiencing a pattern of stimulation of his surface neurons that lie in its stimulus meaning; but one should be careful to remember that a stimulus-true sentence is not necessarily true *simpliciter*. If you show me a life-like replica of a duck, the sentence, "that's a duck", may be stimulus-true for me, but it isn't true. A sentence is defined to be an *observation sentence* for a community if it is an occasion sentence (one whose truth-value is regarded as varying with time and place, although this is not the Quinian definition) and it has the *same* stimulus meaning for all speakers. Thus "He is a bachelor" is not an observation sentence, since different stimulations will prompt you to assent to it than will prompt me (we know different people); but "that's a duck" is (nearly enough) an observation sentence. Observe that the criterion is supposed to be entirely physicalistic. The key idea is that observation sentences are distinguished among occasion sentences by being keyed to the same stimulations *intersubjectively*.

Mach held that talk of unobservables, including (for him) material objects, is justified only for reasons of "economy of thought". The business of science is *predicting regularities in our sensations*; we introduce "objects" other than sensations only as needed to get theories which neatly predict such regularities.

In a recent paper, Quine comes close to a "physicalized" version of

Mach's view. Discussing the question, whether there is more than one correct "system of the world", he gives his criteria for such a system: (1) it must predict a certain number of stimulus-true observation sentences;¹³ (2) it must be finitely axiomatized; (3) it must contain nothing unnecessary to the purpose of predicting stimulus-true observation sentences and conditionals. In the terminology Quine introduces in this paper, the theory-formulation must be a "tight fit"¹⁴ over the relevant set of stimulus-true observation conditionals. (This is a formalized version of Mach's "economy of thought".)

If this were all of Quine's doctrine, there would be no problem. It is reconciling what Quine says here with what Quine says elsewhere that is difficult and confusing. I am *not* claiming that it is impossible however; a lot, if not all, of what Quine says *can* be reconciled. What I claim is that Quine's position is much more complicated than is generally realized.

For example, what is the *status* of Quine's ideal "systems of the world"? It is tempting to characterize the sentences in one of Quine's ideal "theory formulations" as *truths* (relative to that language and that choice of a formulation from among the equivalent-but-incompatible-at-face-value formulations¹⁵ of what Quine would regard as the *same* theory) and as *all* the truths (relative to the same choice of language and formulation), but this would conflict with *bivalence*, the principle that *every* sentence, in the ideal scientific language Quine envisages, is true or false.

To spell this out: Quine's ideal systems of the world are *finitely axiomatizable theories*, and contain standard mathematics. Thus Gödel's celebrated result applies to them: there are sentences in them which are neither provable nor refutable on the basis of the system. If being *true* were just being a theorem in the system, such sentences would be neither true nor false, since neither they nor their negations are theorems. But Quine holds to bivalence.¹⁶

If Quine were a metaphysical realist there would again be no problem: the ideal system would contain everything that could be *justified* (from a very idealized point of view, assuming knowledge of all observations that *could* be made, and logical omniscience); but, Quine could say, the undecidable sentences are still determinately true or false – only we can't tell which. But the rejection of metaphysical realism, of the whole picture of a determinate "copying" relation between words and a noumenal world, is at the heart of Quine's

philosophy. And, as we shall see in the next section, "justification" is a notion Quine is leery of. So what is he up to?

I hazard the following interpretation: bivalence has *two* meanings for Quine: a "first order" meaning, a meaning as viewed *within* the system of science (including its Tarskian metalanguage) and a "second order" meaning, a meaning as viewed by the philosopher. In effect, I am claiming that Quine too allows himself a "transcendental" standpoint which is different from the "naive" standpoint that we get by just taking the system at face value. (I am not claiming that this is *inconsistent* however; some philosophers feel that such a move is *always* an inconsistency, but taking this line would preclude using *any* notion in science which one would explain away as a useful fiction in one's commentary on one's first order practice. There was an inconsistency in the case of the methodological solipsist, because he claimed his first order system reconstructed the *only* way he could understand the notion of an other mind; if he withdraws that claim, then his position becomes perfectly consistent; it merely loses all philosophical interest.)

From *within* the first order system, "*p* is true or *p* is false" is simply true; a derivable consequence of the Tarskian truth definition, given standard propositional calculus. From *outside*, from the meta-metalinguistic point of view Quine occupies, there is no unique "world", no unique "intended model". Only *structure* matters; every model of the ideal system (I assume there is just one ideal theory, and we have fixed a formulation) is an intended model. Statements that are provable are true in *all* intended models; undecidable statements are true or false in each intended model, but not *stably* true or false. Their truth value varies from model to model.

If *this* is Quine's view, however, then there is still a problem. For Quine, what the philosopher says from the "transcendental" standpoint is subject to the same methodological rules that govern ordinary first order scientific work. Even mathematics is subject to the same rules. Mathematical truths, too, are to be certified as such by showing they are theorems in a system which we need to predict sensations (or rather, stimulus-true observation conditionals), given the physics which we are constructing as we construct the mathematics. More precisely, the *whole system of knowledge* is justified as a *whole* by its utility in predicting observations. Quine emphasizes that there is no room in this view for a special status for philosophical utterances.

There is no “first philosophy” above or apart from science, as he puts it.

Consider, now, the statement:

A statement is *rightly assertible* (true in all models) just in case it is a theorem of the relevant “finite formulation”, and that formulation is a “tight fit” over the appropriate set of stimulus-true observation conditionals.

This statement, like most philosophical statements, does not imply *any* observation-conditionals, either by itself or in conjunction with physics, chemistry, biology, etc. Whether we say that some statements which are undecidable in the system are really rightly assertible or deny it does not have any effects (that one can foresee) on prediction. Thus, *this statement cannot* itself be rightly assertible. In short, *this* reconstruction of Quine’s positivism makes it *self-refuting*.

The difficulty, which is faced by all versions of positivism, is that positivist exclusion principles are always self-referentially inconsistent. In short, *positivism produced a conception of rationality so narrow as to exclude the very activity of producing that conception*. (Of course, it also excluded a great many other kinds of rational activity.) The problem is essentially sharp for Quine, because of his explicit rejection of the analytic/synthetic distinction, his rejection of a special status for philosophy, etc.

It may be, also, that I have just got Quine wrong. Quine would perhaps reject the notions of “right assertibility”, “intended model”, and so on. But then I just don’t know what to make of this strain in Quine’s thought.

‘EPISTEMOLOGY NATURALIZED’

‘Epistemology Naturalized’ takes a very different tack. “Justification” has failed. (Quine considers the notion only in its strong “Cartesian” setting, which is one of the things that makes this paper puzzling.) Hume taught us we *can’t* justify our knowledge claims (in a foundational way). Conceptual reduction has also failed (Quine reviews the failure of phenomenism as represented by Carnap’s attempt in the *Logische Aufbau*). So, Quine urges, let us give up epistemology and “settle for psychology”.

Taken at face value, Quine's position is sheer epistemological Eliminationism: we should just *abandon* the notions of justification, good reason, warranted assertion, etc., and *reconstrue* the notions of "evidence" (so that the "evidence" becomes the sensory stimulations that *cause us* to have the scientific beliefs we have). In conversation, however, Quine has repeatedly said that he didn't mean to "rule out the normative"; and this is consistent with his recent interest in such notions as the notion of a "tight fit" (an economical finitely axiomatized system for predicting observations).

Moreover, the expression "naturalized epistemology" is being used today by a number of philosophers who explicitly consider themselves to *be* doing normative epistemology, or at least methodology. But the paper, 'Epistemology Naturalized', really does rule all that out. So it's all *extremely* puzzling.

One way to reconcile the conflicting impulses that one sees at work here might be to replace justification theory by reliability theory in the sense of Goldman; instead of saying that a belief is justified if it is arrived at by a reliable method, one might say that the notion of justification should be *replaced* by the notion of a verdict's being the product of a reliable method. This is an *Eliminationist* line in that it does not try to reconstruct or analyze the traditional notion; that was an intuitive notion that we now perceive to have been defective from the start, such a philosopher might say. Instead, he proposes a *better* notion (by his lights).

While some philosophers would, perhaps, move in this direction, Quine would not for a reason already given: Quine rejects metaphysical realism, and the notion of reliability presupposes the notion of *truth*. Truth is, to be sure, an acceptable notion for Quine, if defined à la Tarski, but so defined, it cannot serve as the primitive notion of epistemology or of methodology. For Tarski simply defines "true" so that "p is true" will come out equivalent to "p"; so that, to cite the famous example, "Snow is white" is *true* will come out equivalent to "Snow is white". What the procedure does is to define "true" so that saying that a statement is true is equivalent to *assenting* to the statement; truth, as defined by Tarski is not a *property* of statements at all, but a syncategorematic notion which enables us to "ascend semantically", i.e., to talk about sentences instead of about objects.¹⁷

I will assent to "p is true" whenever I assent to p; therefore, I will

accept a method as reliable whenever it *yields verdicts I would accept*. I believe that, in fact, this is what the “normative” becomes for Quine: the search for methods that yield verdicts that one oneself would accept.

WHY WE CAN'T ELIMINATE THE NORMATIVE

I shall have to leave Quine's views with these unsatisfactory remarks. But why not take a full blown Eliminationist line? Why *not* eliminate the normative from our conceptual vocabulary? Could it be a superstition that there is such a thing as reason?

If one abandons the notions of justification, rational acceptability, warranted assertibility, right assertibility, and the like, completely, then “true” goes as well, except as a mere device for “semantic ascent”, that is, a mere mechanism for switching from one level of language to another. The mere introduction of a Tarskian truth-predicate cannot define for a language any notion of *rightness* that was not already defined. To reject the notions of justification and right assertibility while *keeping a metaphysical realist notion of truth* would, on the other hand, not only be peculiar (what ground could there be for regarding truth, in the “correspondence” sense, as *clearer* than right assertibility?), but incoherent; for the notions the naturalistic metaphysician uses to explain truth and reference, e.g., the notion of causality (explanation), and the notion of the *appropriate type* of causal chain depend on notions which presuppose the notion of reasonableness.

But if *all* notions of rightness, both epistemic and (metaphysically) realist are eliminated, then what are our statements but noise-makings? What are our thoughts but *mere* subvocalizations? The elimination of the normative is attempted mental suicide.

The notions, “verdict I accept” and “method that leads to verdicts I accept” are of little help. If the *only* kind of rightness any statement has that I can understand is “being arrived at by a method which yields verdicts *I accept*”, then I am committed to a solipsism of the present moment. To solipsism, because this *is* a methodologically solipsist substitute for assertibility (“verdicts *I accept*”), and we saw before that the methodological solipsist is only consistent if he is a real solipsist. And to solipsism of the present moment because this is

a *tensed* notion (a substitute for warranted assertibility at *a time*, not for assertibility in the best conditions); and if the *only* kind of rightness my present “subvocalizations” have is *present* assertibility (however defined); if there is no notion of a *limit* verdict, however fuzzy; then there is no sense in which my “subvocalizations” are *about* anything that goes beyond the present moment. (Even the thought “there is a future” is “right” only in the sense of being *assertible at the present moment*, in such a view.)

One could try to overcome this last defect by introducing the notion of “a verdict I would accept *in the long run*”, but this would at once involve one with the use of counterfactuals, and with such notions as “similarity of possible worlds”. But it is pointless to make further efforts in this direction. Why should we expend our mental energy in convincing ourselves that we aren’t thinkers, that our thoughts aren’t really *about* anything, noumenal or phenomenal, that there is *no* sense in which any thought is *right* or *wrong* (including the thought that no thought is right or wrong) beyond being the verdict of the moment, and so on? This is a self-refuting enterprise if there ever was one! Let us recognize that one of our fundamental self-conceptualizations, one of our fundamental “self-descriptions”, in Rorty’s phrase, is that we are *thinkers*, and that as thinkers we are committed to there being *some* kind of truth, some kind of correctness which is substantial and not merely “disquotational”. That means that there is no eliminating the normative.

If there is no eliminating the normative, and no possibility of reducing the normative to our favorite science, be it biology, anthropology, neurology, physics, or whatever, then where are we? We might try for a grand theory of the normative in its *own* terms, a formal epistemology, but that project seems decidedly over-ambitious. In the meantime, there is a great deal of philosophical work to be done, and it will be done with fewer errors if we free ourselves of the reductionist and historicist hang-ups that have marred so much recent philosophy. If reason is both transcendent and immanent, then philosophy, as culture-bound reflection and argument about eternal questions, is both in time and eternity. We don’t have an Archimedean point; we always speak the language of a time and place; but the rightness and wrongness of what we say is not just for a time and a place.

NOTES

* Delivered at the University of California, Berkeley, on April 30, 1981. This was the second of two Howison Lectures on 'The Transcendence of Reason'. The first one has appeared in *Synthese* 51 (1982), 141-167.

¹ See his 'Universals' for an illuminating account of this tension.

² See my *Reason, Truth and History*.

³ I chose brown because brown is not a spectral color. But the point also applies to spectral colors: if being a color were purely a matter of reflecting light of a certain wavelength, then the objects we see would change color a number of times a day (and would all be black in total darkness). Color depends on background conditions, edge effects, reflectancy, relations to amount of light etc. Given a description of all of these would only define *perceived* color; to define the "real" color of an object one also needs a notion of "standard conditions": traditional philosophers would have said that the color of a red object is a power (a disposition) to look red to normal observers under normal conditions. This, however, requires a counterfactual conditional (whenever the object is *not* in normal conditions) and we saw in the previous lecture that the attempt to define counterfactuals in "physical" terms has failed. What makes color terms physically undefinable is not that color is subjective but that it is *subjunctive*. The common idea that there is some one molecular structure (or whatever) common to all objects which look red "under normal conditions" has no foundation: consider the difference between the physical structure of a red star and a red book (and the difference in what we count as "normal conditions" in the two cases).

⁴ This argument appears in Firth's Presidential Address to the Eastern Division of the American Philosophical Association (Dec. 29, 1981), titled 'Epistemic Merit, Intrinsic and Instrumental'. Firth does not specifically refer to evolutionary epistemology, but rather to "epistemic utilitarianism"; however, his argument applies as well to evolutionary epistemology of the kind I describe.

⁵ See his 'What is Justified Belief'.

⁶ See 'A Plea for Excuses'.

⁷ See Rorty's 'Pragmatism, Relativism and Irrationalism'.

⁸ *Philosophy and the Mirror of Nature*.

⁹ In *The Claim of Reason*.

¹⁰ In *Reason, Truth and History*.

¹¹ *The Claim of Reason*, p. 125.

¹² W.V. Quine, *Ontological Relativity and Other Essays*.

¹³ Quine actually requires that a "system of the world" predict that certain "pegged observation sentences" be true. I have oversimplified in the text by writing "observation sentence" for "pegged observation sentence". Also "the stimulus meaning" of an observation sentence includes a specification of conditions under which the speaker *dissents*, as well as the conditions under which he *assents*. The details are in the paper 'On Empirically Equivalent Systems of the World'.

¹⁴ A theory is a "tight fit" if it is interpretable in *every* axiomatizable theory which implies the observation conditionals (conditionals whose antecedent and consequent are pegged observation sentences) in question in a way that holds the pegged observation sentences fixed. To my knowledge, no proof exists that a "tight fit" even exists, apart from the trivial case in which the observation conditionals can be axiomatized without going outside of the observation vocabulary.

¹⁵ Two theories (in the usual sense) are regarded as "formulations" of the *same* theory by Quine if each is interpretable in the other in a way that holds the pegged observation sentences fixed.

¹⁶ See 'What Price Bivalence'.

¹⁷ Quine himself puts this succinctly. "Whatever we affirm, after all, we affirm as a statement within our aggregate theory of nature as we now see it; and to call a statement true is just to reaffirm it". ('Empirically Equivalent Systems of the World', p. 327).

REFERENCES

- Austin, J.: 'A Plea for Excuses', in *Philosophical Papers*, 2nd Edition, Oxford, 1970, pp. 175-204.
- Cavell, S.: *The Claim of Reason*, Oxford, the Clarendon Press, 1979.
- Firth, R.: 'Epistemic Merit, Intrinsic and Instrumental', forthcoming in the volume of *Proceedings and Addresses of the American Philosophical Association* that will contain the Presidential Addresses from 1981-82. (Delivered to the Eastern Division of the A.P.A., Dec. 29, 1980).
- Goldman, A.: 'What is Justified Belief', forthcoming in George S. Pappas (ed.) *Justification and Knowledge*.
- Putnam, H.: *Reason, Truth and History*, Cambridge University Press, 1981.
- Quine, W. V.: 'Ontological Relativity', in *Ontological Relativity and Other Essays*, Columbia University Press, 1969.
- Quine, W. V.: 'On Empirically Equivalent Systems of The World', *Erkenntnis* 9 (1975), pp. 313-328.
- Quine, W. V.: 'What Price Bivalence', *The Journal of Philosophy*, 78 (1981), pp. 90-95.
- Rorty, R.: *Philosophy and the Mirror of Nature*, Princeton University Press, 1979.
- Rorty, R.: 'Pragmatism, Relativism and Irrationalism', Presidential Address to the Eastern Division of the American Philosophical Association (Dec. 29, 1979), in *Proceedings and Addresses of the American Philosophical Association* 53 (1980).
- Strawson, P.: 'Universals', *Midwest Studies in Philosophy*, Vol. IV, University of Minnesota Press, 1979, pp. 3-10.

THE RELATION BETWEEN EPISTEMOLOGY AND PSYCHOLOGY

In the wake of Frege's attack on psychologism and the subsequent influence of Logical Positivism, psychological considerations in philosophy came to be viewed with suspicion. Philosophical questions, especially epistemological ones, were viewed as 'logical' questions, and logic was sharply separated from psychology. Various efforts have been made of late to reconnect epistemology with psychology. But there is little agreement about how such connections should be made, and doubts about the place of psychology within epistemology are still much in evidence. It therefore remains to be clarified just how such links should be established, and what impact they would have on the direction of epistemology.

One reason people say very different things about the bearing of psychology on epistemology is their different conceptions of the aims of epistemology. Let us begin with a brief delineation of three such conceptions (neither exhaustive nor mutually exclusive): (A) *descriptive epistemology*, (B) *analytical epistemology*, and (C) *normative epistemology*. If the first of these conceptions is adopted, it should not be controversial that psychology has a vital role to play. By contrast, if the third conception is adopted, it may seem highly doubtful that psychology is relevant at all. I shall briefly survey the role that psychology might have in descriptive and analytical epistemology, and then focus more detailed attention on normative epistemology. I shall argue that psychology has an important contribution to make *even* to normative epistemology.

1. DESCRIPTIVE EPISTEMOLOGY

Textbook definitions frequently include among the tasks of epistemology the identification of "sources" of knowledge, that is, ways in which knowledge can be acquired. This strongly suggests that epistemology is concerned with the psychological processes of knowledge-acquisition, or more generally with belief acquisition. Such an interpretation is confirmed by the historical literature, which is replete with

descriptions and classifications of mental faculties and endowments, processes and contents, acts and operations. A sampling of this literature would reveal the following examples of mental or cognitive classification, all relevant to knowledge-acquisition.

Cognitive faculties: The senses, reason, memory, intuition, the active and the passive intellect, the understanding, the imagination, and the will.

Cognitive acts or processes: Sensing, judging, doubting, imagining, conceiving, intuiting, recollecting, introspecting, comparing, compounding, distinguishing, abstracting, associating, synthesizing, schematizing, and generalizing.

Cognitive contents: Ideas, impressions, concepts, and categories.

Classifications of contents, e.g., ideas, either in terms of their intrinsic character or their origin: Simple, complex, clear, confused, innate, acquired, forceful, lively, vivacious.

Analytic philosophers have tended to criticize or minimize the importance of philosophical description in the historical writings. It is common, for example, to regard the mentalistic dissections of the British Empiricists as symptomatic of a confusion, a failure to draw a proper distinction between psychological questions and epistemological questions. D. W. Hamlyn (1967, p. 9) writes: "Epistemology differs from psychology in that it is not concerned with why men hold the beliefs they do or with the ways in which they come to hold them." This remark, though characteristic of the analytic approach, is anomalous in its context. It occurs in an introductory section of Hamlyn's encyclopedia article on the history of epistemology, and the ensuing bulk of the article makes it perfectly plain that, historically speaking, epistemologists most certainly *were* interested in "why men hold the beliefs that they do", or "the ways in which they come to hold them". The legacy of Logical Positivism and the Ordinary Language movement was a bifurcation of epistemology and psychology. The latter may be concerned with the origin of ideas; the former should be concerned with their validity, justification, or logical cogency. But even if one accepts this proposed distinction, it cannot be denied that the *historical* philosophers had a thoroughgoing interest in mental or cognitive processes. *Their* conception of epistemology was at least *partly* descriptive, and this descriptive component consisted primarily of the description of mental acts and operations.

Moreover, this descriptive approach to the epistemology is not confined to pre-20th-century writers. In the late 1960's, W. V. Quine (1969 and 1975) began advocating an enterprise called "naturalistic epistemology", the project of describing and explaining how human beings come to hold their theory of the world, and how it happens that this theory is so successful. Naturalistic epistemology is expressly classified as a branch of empirical psychology, and Quine's learning theoretic approach, as presented, for example, in *The Roots of Reference* (1974), is some form of behavioristic psychology.

Jean Piaget is another influential recent writer for whom epistemology is an empirical, descriptive subject, partly psychological in nature. Piaget describes his brand of epistemology – "genetic epistemology" – as "the theory of scientific knowledge founded on the development of this knowledge." "Genetic epistemology is the study of successive states of a science as a function of its development". It is "the study of the mechanisms of the growth of knowledge", including, as a proper part, the growth of psychological or cognitive mechanisms in the individual.¹

A third recent proponent of epistemology as an empirical, descriptive discipline is the psychologist Donald Campbell. In his earlier work Campbell advocated an evolutionary epistemology, and his recent William James Lectures characterize his conception as "Descriptive Epistemology: Psychological, Sociological, and Evolutionary."²

Indeed, certain components of the heritage of historical epistemology have been taken over by psychology and Artificial Intelligence. Many concerns of the classical philosophers about mental processes are explored by cognitive psychology, developmental psychology, social psychology, psycholinguistics, and AI. While these investigators do not standardly apply the label 'epistemology' to their enterprises, some members of the AI community have even appropriated this term, and psychologists are increasingly aware of the close ties between their investigations and epistemology (Goldstein and Papert, 1977; Nisbett and Ross, 1980).

It should not be denied, then, that descriptive epistemology is one conception of the subject, and psychology is an essential element in this conception. (Psychology would not exhaust descriptive epistemology, since epistemology has important *social* as well as *individual* dimensions. Descriptive social epistemology would include such disciplines as the sociology of science and knowledge, the history of science, and cultural anthropology. But our focus here is on *individual* epis-

temology.) At the same time, it is obvious that the conception of epistemology as a purely factual or descriptive subject is a minority view, at least within philosophical circles. Even if we concentrate on historical writers, concern with the genesis of belief was arguably ancillary to their main inquiry, i.e., is skepticism warranted, or can valid knowledge be attained? Furthermore, whatever motivated the classical philosophers, there is a widespread view in this century that there are distinctively philosophical questions about knowledge, rationality, and methodology – questions that cannot be settled by empirical description. These distinctively philosophical questions deserve to be set aside as the special domain of epistemology quite separate from empirical description.

For these reasons, many philosophers would dismiss “naturalistic epistemology”, “genetic epistemology”, or “evolutionary epistemology” as not proper conceptions of epistemology at all. I regard this response as too extreme. There is no need to preempt the term “epistemology”, to restrict its use to a single intellectual enterprise. Nonetheless, I would agree that on some of the most important and influential conceptions of epistemology, the questions of epistemology do not seem to be answerable by purely descriptive analyses of psychological processes. We need to explore what these questions are, and how they should be addressed.

2. ANALYTICAL EPISTEMOLOGY

According to analytical epistemology, the main task of the subject is to *analyze* key epistemological concepts or terms, such as *knowledge*, *rationality*, *justification*, and the like. “Analysis” consists of explicating or defining the concepts or terms in question, and perhaps in making precise or refining them where they are vague or confused.³ This is standardly regarded as a distinctively philosophical activity, not an empirical or experimental one.

However, even the “analysis” of epistemological concepts brings important mentalistic or psychological subject matter into the picture. It is widely agreed, for example, that the concept of knowledge involves the concept of belief, clearly a mentalistic notion. Furthermore, if some version of a causal theory of knowledge is correct, mentalistic or psychological subject matter enters into the analysis of knowledge in a

more pervasive fashion. Consider, for example, an approach I have advocated in a recent paper (Goldman, 1979; cf. 1976). On this approach, justified belief is a necessary condition for knowing, and a causal account of justified belief is advanced. More specifically, the theory of justified belief defended there is "*Historical Reliabilism*". The theory suggests that there are various types of belief-forming processes, some of which are more *reliable* than others; that is, they tend to lead to truth a greater proportion of the time. Complications aside, Historical Reliabilism says that a particular belief is justified just in case its historical ancestry consists of cognitive processes that are generally reliable, i.e., lead to truth in a sufficiently high proportion of cases. Such a theory might be called a "naturalistic" theory of knowledge, because it locates the source of justification in "natural" facts or processes, i.e., in certain psychological processes. But this kind of "naturalistic epistemology" should be distinguished from Quine's brand, since it appears in the guise of "conceptual analysis", which Quine would reject.

The fact that mental or psychological processes are sources of knowledge and justified belief does not imply that epistemology must appeal to the empirical science of psychology. If epistemology is a branch of conceptual analysis, it is only interested in the *ordinary* person's *concept* of knowledge and *concept* of justified belief. If these concepts presuppose categories and theories of the mental, an adequate epistemological theory will identify these mentalistic commitments in the ordinary person's framework. It is irrelevant, however, whether these commonsense mentalistic commitments are correct or incorrect, accurate or inaccurate. Empirical psychology may look askance at the ordinary person's views of belief-forming processes, but that is irrelevant to a mere "analysis" of the ordinary concept of knowledge. Thus, if "analysis" is the aim of epistemology, the science of psychology has no obvious entree into the epistemological enterprise.

It might be argued, though, that the very activity of conceptual analysis is a species of psychological investigation, or applied psychology. Arguably, a concept is some sort of mental representation.⁴ To analyze a concept is to identify or characterize a mental representation; and such an activity properly belongs within the sphere of psychology. A similar argument holds if analysis is construed as concerned with meanings of words. Words in public languages ultimately get their meanings from mental representations that underlie their production and comprehension.⁵ So analysis of word meanings

would still be ultimately concerned with mental representations, and could still be construed as a province of applied psychology, or applied psycholinguistics.

I have considerable sympathy with this position, though this is not the place to argue for it in detail. However, if this view is right, *all* philosophical analysis, not just *epistemological* analysis, becomes the province of psychology. There will be no *distinctive* connections of psychology with epistemology, as compared with other branches of philosophy. My interest here, however, is to explore the *special* connections between epistemology and psychology. So let us set aside the question of whether the activity of "analysis" is fundamentally a psychological activity.

Have we done justice to the analytical conception of epistemology? Epistemology, many analytic philosophers would insist, is not merely a matter of defining certain words, or even refining and precising their use. It is also a matter of determining *proper* methods of inquiry, of identifying *legitimate* procedures for forming beliefs or other doxastic attitudes. Epistemology, even analytical epistemology, is interested in specifying *rules* or *principles* that prescribe, permit, or prohibit various intellectual attitudes or strategies. Furthermore, the specification of such rules or principles cannot fall within the province of psychology. Psychology is a factual or positive science, not a normative discipline. The selection of rules and principles belongs to a normative discipline. Hence, empirical psychology cannot supplant epistemology, and indeed has no *relevance* at all to *this* task of epistemology. So goes an important argument for detaching psychology from epistemology, at least from one very central component of epistemology.

3. NORMATIVE EPISTEMOLOGY

The first-mentioned approach to epistemology – the rule-giving or principle-stating approach – is the third conception introduced at the outset: normative epistemology. It is exemplified by epistemologists of various persuasions, such as R. M. Chisholm (1966 and 1977) on the one hand and Bayesians on the other.⁶ In the remainder of the paper I shall concentrate my attention on normative epistemology.

As we have seen, epistemology is often concerned with rationality and irrationality, with justification and the absence of justification, with

warranted and unwarranted belief. Now 'rational' and 'irrational', 'justified' and 'unjustified', 'warranted' and 'unwarranted' are all *evaluative* or *normative* expressions. They are used to appraise or grade intellectual acts or strategies along the epistemic (as opposed to the ethical or aesthetic) dimension. To say that a belief is *unwarranted* is to say that, epistemically speaking, the cognizer *ought not* have that belief. To say that a belief is *warranted*, on the other hand, is to say that the cognizer is epistemically *entitled* to the belief, that he has an epistemic *right* to it, or that it is *permitted* or *licensed* from an epistemic point of view. All of these terms – 'entitled', 'right', 'permitted', 'licensed' – have a deontic flavor, and this flavor might be clarified by linking terms of epistemic appraisal with *rules* or *principles*, of the sort alluded to a moment ago.

This approach may be motivated by an example. Suppose you arrive at a conference and join a number of unfamiliar people at dinner. Each person introduces himself, the food arrives, and all commence eating. At this juncture, Jones utters the sounds PLEEZ PASS TH' SAHLT. You form a *belief* in the proposition, *Jones wants the salt*. Is this belief of yours *justified*? Assume that immediately prior to forming this belief, you had the following (justified) beliefs. (i) You believed that Jones uttered the sounds: PLEEZ PASS TH' SAHLT. (ii) You believed that Jones is an English-speaker (since he introduced himself with appropriate English utterances). (iii) You believed that the phonetic sequence, PLEEZ PASS TH' SAHLT, is an appropriate way in English of asking for the salt. (iv) You believed that there is meat on Jones' dish, and that most people prefer their meat salted. (v) You believed that there is salt on the table, and that Jones can see that there is. Given all these antecedent beliefs, most of us would be inclined to say that your belief in the proposition, *Jones wants the salt*, is a *justified* belief. One way of understanding this inclination is to say that it stems from a tacit commitment to certain rules or principles, rules that say what *transitions* it is appropriate to make from certain cognitive states (e.g., belief-states) to other cognitive states. Most of us would agree that it is appropriate to move *from* a state of believing that Jones uttered the indicated sounds, that Jones is an English-speaker, that this sequence of sounds is an appropriate way in English of asking for the salt, etc., *to* a state of believing the proposition that Jones wants the salt. In short, most of us would regard the target belief as justified because we view the indicated cognitive-state-transition as a transition that *conforms*

with correct epistemic rules. This illustrates the general idea that the justificational status of a belief (or other doxastic attitude) is determined by its conformity or nonconformity with correct principles of cognitive-state-transition. If a belief can be arrived at from an immediately prior cognitive state in accordance with the right principles, then the person would be *justified* in having that belief, would be *entitled* to that belief. More precisely, the resultant belief would be justified if the transition is appropriate *and* if the beliefs in the prior cognitive state are themselves justified, which depends on whether *they* were generated in conformity with correct epistemic rules. In this way, the justificational status of a belief is a function of its *history*, just as Historical Reliabilism maintains.

Although most people would say, in the salt-passing case, that you are justified in believing that Jones wants salt, someone of a skeptical turn of mind might think otherwise. After all, he might argue, the conclusion that Jones wants the salt is underdetermined by the evidence; there are many alternative possible explanations of why Jones uttered PLEEZ PASS TH' SAHLT. Perhaps you are justified in assigning this hypothesis a reasonably high probability, but you aren't justified in *believing* it. Apparently, this skeptic differs from most of us concerning the correct epistemic rules. In his view, the correct rules don't license *belief* in the indicated conclusion. The skeptic agrees, though, that justificational status, either of a belief or a subjective probability, depends on conformity with the correct rules. His only disagreement concerns *which* rules are correct. Thus, the proposed link between justifiedness and conformity-with-correct-rules still stands. Indeed, it is a neutral and generally acceptable idea. It may be formulated as the following *meta-epistemological* principle.

Principle I: A person's doxastic state D (at time t) is justified if and only if D results from a history of cognitive-state transitions that conform with (are permitted by) the correct epistemic rules.

Although I shall later make a relatively minor change in this principle, I shall appeal to it on several occasions to test the correctness of various candidate rules. In other words, it is a high-level principle of our meta-theory.

Clearly, however, Principle I does not directly specify the correct epistemic rules. What, then, *does* determine which rules are correct? What is needed are *right-making characteristics* for epistemic rules.

What kinds of characteristics are there whose possession would *make* a given epistemic rule (or set of rules) the *right* rule? Or what characteristics would make one rule (or set of rules) *better than* an alternative rule? If the right-making characteristics can be settled upon, the question of *which* rules have those features, or have more of them than other rules, could be investigated. But it is hopeless to try to decide which rules are better than others until we first settle the question of what would *make* one candidate rule (or set of rules) better than another.

A familiar and plausible view is that the aim of inquiry is to attain truths and avoid errors, i.e., to *believe* truths and avoid believing falsehoods. This suggests the following right-making characteristic for epistemic rules.

Principle II: *An epistemic rule R is right, or correct, if and only if R belongs to a complete set of rules R* such that*

- (A) *human beings can conform with R*, and*
- (B) *conforming with R* would (in the long run) promote more true belief and error-avoidance than conforming with any other complete set of rules with which human beings can conform.*

There are several things unsatisfactory about Principle II. For one thing, it is vague: it doesn't explain how truth-acquisition is to be weighed as compared with error-avoidance. A liberal rule might lead to more true beliefs but also more errors than a cautious rule, and Principle II doesn't say which is better. Second, it places too much emphasis on the *number* of truths acquired, rather than on their quality or importance (either their intrinsic importance, or their importance relative to the cognizer's project). It favors the "hedgehog" over the "fox". Despite these points, the principle will do for present purposes. It serves here as illustrative of the *kind* of right-making approach I favor. I do not endorse it in detail, nor shall my main arguments depend on (all of) its details.

A few comments about it are in order, however. First, the right-making theory expressed by Principle II is a species of epistemological *Consequentialism*, analogous to some varieties of ethical Consequentialism. I think that Consequentialism is needed in epistemology, but I cannot argue for that fully in this paper. Second, since this form of Consequentialism stresses true-belief acquisition, it is close in spirit to

the Reliabilist theory discussed earlier. When Principle II is conjoined with Principle I, a belief turns out to be justified just in case it is produced by a series of transitions of types that are maximally conducive to getting truths and avoiding falsehoods. This is quite close to Historical Reliabilism.

Third, Principle II makes the rightness, or correctness, of a rule independent of what people currently take the right set of rules to be. This is a desirable feature, to my mind, for it safeguards epistemological objectivism. Objectivism is ensured since the rightness of a rule doesn't depend on people's opinions or intuitions about its rightness. Given Principle II, there is a "fact of the matter" concerning which rules are right (assuming, at any rate, that there is a fact of the matter concerning the relevant propensities, or subjunctives). This fact of the matter transcends whatever opinions people may have about which rules are best. It follows, moreover, that no "conservative" standpoint is warranted, automatically endorsing our present intellectual practice (or "language game"). Our present practice, or the rules to which we presently subscribe, may be sub-optimal; there may be room for improvement. Indeed, Principle II provides a *standard* by which to judge whether revisions are in order, and if so, what revisions they should be. Precisely this sort of standard has presumably been employed in the past when canons of scientific methodology have been advanced as superior to pre-scientific thinking.

A fourth comment concerns the nature of the rules. Are they supposed to be *given* to cognizers, who are expected to *apply* them in making cognitive choices? Are they, in other words, cognitive *decision-making* rules? Or are they merely rules for *theoretical* appraisal, i.e., rules an external observer might use to assess the propriety or impropriety of a person's cognitive procedures? I am thinking here only of the second function. The rules may be very complicated and difficult to understand, so that few cognizers – especially children or untutored adults – would be able to appeal to these rules to guide their intellectual conduct. These characteristics would make the rules unsuitable as "decision guides", but they might still be suitable for the theoretical purpose. Only this theoretical purpose is in question here. When Principles I and II speak of "conformity" with a rule, then, they simply mean that the state-transitions "fit" the rule, or "accord" with it. They don't mean that the cognizer "follows" the rule, or uses it as a

“decision-guide”. A person can have justified beliefs if he forms beliefs *in accordance* with correct rules, even if he doesn’t *know* or *appeal* to these rules.

A final question concerning Principle II pertains to its optimization requirement. The principle requires a right rule to belong to a *uniquely* optimal set of rules, but it isn’t clear that there is a unique optimum. Furthermore, is it obvious that a right rule must belong to an optimal set, rather than one of several equally best sets of rules? Or why not adopt a “satisficing” approach, in which any “adequately” truth-conducive set of rules will be deemed right? The problem is what to do with two or more equally optimal, or “satisfactory”, sets that give *conflicting* instructions. Suppose one best set of rules permits a certain belief under specified conditions and another best set of rules forbids it. If a cognizer goes ahead and believes in those circumstances, is his belief justified or not? How can Principle I be applied if there are conflicting, equally right, sets of rules? Since I cannot answer this question, I shall assume that there is a uniquely optimal set, and I shall stick with Principle II which identifies that set as the right set of rules.

Let us work, then, with Principle II. Given this principle, how should we go about trying to select the particular set of rules that are right? What *kinds* of information are relevant to this selection process? Specifically, is information from empirical psychology relevant? I remarked earlier that the aim of the paper is to show the relevance of empirical psychology even to normative epistemology. But no use has been made of psychology in arguing for the meta-epistemological principles that have been adduced. So the only place for empirical psychology within normative epistemology would be in the process of comparing alternative sets of candidate rules. But how and why should empirical psychology enter there?

One answer is that the candidate rules must be rules with which human beings can conform, as Principle II indicates. Clearly, psychology is relevant to this issue. Human capacity or incapacity to conform with various sets of rules should be determined by psychology. I shall return to this point in Section 6. But psychology is also relevant for other reasons.

Many epistemologists have held that among the needed epistemic rules are rules for perceptual belief, and rules for belief based on memory. Rules of the first sort might say, roughly, that if a person is “appeared to” in a certain way, then he is entitled to believe such-and-

such a proposition. Rules of the second sort might say, roughly, that if a person *seems* to remember witnessing such-and-such an event, he is entitled to believe that he really did witness such an event. It is plausible to hold that the best rules of either kind should be informed by whatever psychological facts are uncovered about our perceptual and memorial systems. If our senses are subject to illusions of certain sorts, rules for perceptual belief should be tailored to accommodate the likelihood of such illusions. If our memory systems are subject to certain kinds of unreliability, the rules for relying on memory should reflect these facts. For example, many psychologists view the process of remembering as a "constructive" process, in which the cognizer's representation of a past event is constructed from all sorts of material stored in memory, not just material that originates from the target event.⁸ Thus, material that enters memory some time after a witnessed event – e.g., material derived from the verbal comments of an interrogator – can easily distort or influence one's subsequent "construction" of the target event (Loftus, 1979). If this is right, it is plausible to tailor rules for belief based on memory with these facts in mind. These kinds of considerations are implicitly acknowledged by epistemologists when, for example, they reject the epistemic certainty or incorrigibility of perceptual belief, or when they advocate taking prior beliefs into account rather than relying exclusively on "the testimony of the senses". The senses are not to be trusted unreservedly because they are capable of deception, and this capability is a psychological (or physiological) fact.

If these cases are granted, psychology is already established as relevant to the selection of epistemic rules. A third reason why psychology is relevant has to do with the notion of cognitive-state transitions. Since epistemic rules are concerned with cognitive states, the formulation of rules must take into account, as well as possible, the *kinds* of cognitive states a human being can be in. Past epistemology has tended to work with the mental classifications of commonsense, or "folk", psychology, such classifications as "belief" and "certainty". But as I shall suggest in section 6, it is questionable whether epistemology should rest content with folk psychology's repertoire of cognitive states. On the contrary, an optimal normative epistemology should reflect the best possible delineation of available cognitive states, and such a delineation can only come from the psychology of cognition (properly executed).

For the moment, however, let us set this point aside. Assume, for the

sake of argument, that epistemic rules will only employ the terminology of commonsense psychology. How pervasive should psychology be in selecting the content of epistemic rules? Aren't *some* epistemic rules certifiable without appeal to psychology? A *weak* thesis would be that psychology is required to choose rules in certain domains (e.g., perceptual belief and memory belief) but not others. A *strong* thesis would be that psychology is required to certify *all* epistemic rules. I am inclined to endorse the strong thesis. But it can be expected to encounter resistance. Aren't at least some epistemic rules intended to tell us how to *reason* properly? And isn't logic the normative study of reasoning? So shouldn't we expect *some* epistemic rules to be provided by *logic alone*, without appeal to psychology? I take up this view in the next section.

4. CAN LOGIC GIVE US EPISTEMIC RULES?

Philosophers have standardly distinguished between deductive and inductive reasoning, between deductive and inductive logic. It is very controversial, however, whether there is such a thing as inductive logic, so I shall concentrate on deductive logic. I shall accept the contention that deductive logic – formal logic – is a body of necessary truths. The status of these truths is independent of human psychology. Thus, psychologism in logic is rejected. The question before us is whether epistemic rules can be derived from logic alone, without any appeal to facts of human psychology.⁹

We suggested earlier that epistemic rules are rules for cognitive-state transitions. Let us spell this out more fully. First, however, a few preliminary remarks are in order. Epistemic rules need not be concerned exclusively with *mental* acts of cognizers. They may also be concerned with *overt actions*, such as preparing experimental devices, recording observations, orienting one's sensory organs, and the like. Nonetheless, a large part of the domain of epistemic rules are events in the mind-brain, and these will be the focus of our attention. We also remarked earlier that epistemology has a social dimension. Social normative epistemology would be concerned with inter-personal activities and institutions outside the mind. But our present concern is individual epistemology.

A different point concerns the term 'transition'. This suggests that epistemic rules would only be aimed at the *kinematic*, or *diachronic*,

dimensions of cognition. I don't mean to restrict epistemic rules in this way. They would equally be concerned with the *static*, or *synchronous*, dimension. I shall concentrate, though, on the diachronic dimension.

A third comment applies to the term 'cognitive'. I intend this term broadly, to encompass both conscious and nonconscious states, and both propositional and nonpropositional states of mind. Epistemologists have concentrated on propositional attitudes and I shall largely follow this practice. But I leave open the possibility that different mental states may have different sorts of contents, or no contents at all in any literal sense. The basic *framework* of our epistemology is neutral on this issue, though the set of rules eventually adopted should not be neutral.¹⁰

An example of an epistemic rule is the following: "You are permitted to go from a state of believing propositions q_1, q_2, \dots, q_m , at time t , to a state of believing proposition q_n , at the succeeding time t' ". Other rules might feature transitions from subjective probabilities to subjective probabilities. The rules may either specify transitions between *total* cognitive states, or between *partial* cognitive states. Thoroughgoing coherentists would presumably favor the former over the latter, but our framework is neutral on the question. There are three types of epistemic rules, corresponding to the three basic deontic operators. One kind of rule would *prescribe* a transition, or make it *obligatory*. A second would *permit* a transition. And a third would *forbid* or *prohibit* a transition. Given our interest in justifiedness, I shall concentrate on permission rules, since a belief is justified only if it results from *permitted* transitions. (See Principle I.) (If space allowed, we might explore the possibility of rules less rigidly tied to standard deontic categories, rules that 'advise' or 'recommend' rather than prescribe, permit, or prohibit. But let us stick to the more conventional categories.)

Given our characterization of epistemic rules, it should be obvious that logic doesn't *state* or *present* epistemic rules. Logic is completely silent about cognitive states; its subject-matter is wholly different. It is true that logic studies rules called 'rules of inference', and belief-transitions are also often called 'inferences'. But so-called 'rules of inference' in axiomatic systems or natural deductive systems say nothing about beliefs, or other *psychological* states.

Furthermore, while epistemic rules are prescriptive, permissive, or prohibitive in force, the contents of formal logic have no such force. Logic seeks to describe the properties of formulas and various systems of formulas. While the systems logic studies often *contain* rules, logic

itself does not consist of rules. Formal logic has three branches: semantics, proof theory, and recursive function theory. For present purposes, recursive function theory may be ignored. The main contents of semantics are statements to the effect that certain sentences are valid (true in all interpretations), and the main contents of proof theory are statements that certain sentences are provable (or unprovable) in certain axiomatic systems. These kinds of statements are not prescriptive, permissive, or prohibitive in force. They say nothing about what should be done, cognitively speaking, nor what is permissible, nor prohibited. Hence, statements of logic are not *equivalent* to epistemic rules, nor do they logically *imply* the contents of such rules.

Nonetheless, it might be supposed that some epistemic rules can be constructed on the basis of truths of logic, and logic alone. For example, it might be thought that, for any propositions q and r , if q logically implies r (written ' $q \models r$ '), there is a (correct) epistemic rule that permits a transition from a belief in q to be a belief in r . In other words, if q logically implies r , a cognizer is licensed to *infer* r from q , in the psychological sense of 'infer'.

Let us test this proposal. What we have here is not a single rule, but a general *schema* for generating epistemic rules from truths of logic. That is just the sort of thing to be expected if epistemic rules can be constructed from truths of logic. Some notation to represent the proposed schema will be convenient. Let ' P ' be the epistemic permission operator. Let ' B ' be the belief operator, and let ' q ', ' r ', etc., be variables that range over propositions. Let the slash symbol '/' represent cognitive-state transitions, i.e., transitions from a cognitive state at one moment to a cognitive state at a succeeding moment. (Time is treated discretely for the sake of simplicity.) Thus, " $P (\dots / \dots)$ " is a rule that permits a transition from a state represented by the left-hand expression to a state represented by the right-hand expression. We may then write the rule-generating-schema suggested above as follows:

(RGS 1) For any propositions q and r , if $q \models r$, then $P (Bq / Br)$.

The proposed schema generates only permission rules, not obligation rules. If such transitions were obligatory, deductive closure of one's beliefs would be required in a moment of time; one would be required to believe all of the infinitely many propositions implied by prior beliefs. In all likelihood, it is psychologically impossible to conform with such a

requirement; hence no such requirement should be made. The corresponding permission rules have no such liability.

Let us now ask whether RGS 1 is an acceptable schema, a satisfactory way of deriving epistemic rules from truths of logic. Harman (1973, p. 157) has objected to schemas like RGS 1 in the following way. Suppose that q does imply r and you believe q . It doesn't follow that you are entitled to accept r as well. Perhaps what you ought to do is stop believing q . This is, after all, how we sometimes proceed. Upon seeing some unacceptable implications of our prior beliefs, we decide to abandon one or more of the prior beliefs, rather than accept their implications.

This objection is not quite conclusive. One might reply to it as follows. "A proper set of rules would not allow one to form a belief in any proposition that has unacceptable implications. Hence, if you find yourself holding a belief with unacceptable implications, that prior belief must not have been formed (or retained) in accordance with the correct total set of rules. But the rules generated by RGS 1 cannot be held responsible for that outcome. They can only be expected to make *their* contribution to correct epistemic practice. As long as all *other* rules are well-designed, and as long as you conform with them, there won't be a case in which you find yourself believing a proposition with unacceptable implications. Hence, Harman's objection doesn't hold, and RGS 1 is perfectly satisfactory".

It isn't clear whether this reply succeeds. Until we know more about what *other* correct epistemic rules would be like, it isn't obvious whether a cognizer could conform with correct epistemic rules and still have a belief that is later discovered to have unacceptable implications. Furthermore, it isn't clear that epistemic rules should be designed on the assumption that cognizers will always follow them. Perhaps rules should take into account the possibility that prior beliefs may not have been properly formed.¹¹ If so, then Harman's objection would hold, since a badly formed belief may well be discovered to have unacceptable implications, and the right thing to do there is to drop the prior belief, not accept its newly discovered implications. For these reasons, Harman's objection may survive the reply.

However, the argument against RGS 1 need not rest on Harman's objection. Difficulties face it from other quarters. Suppose that Oscar believes some proposition q , which in fact logically implies r , but Oscar doesn't *realize*, i.e., doesn't believe, that it logically implies r . In

particular, suppose that both q and r are complex propositions, and the entailment is difficult to recognize. For example, q might be ' $(x)(y)(Fx \supset Gy)$ ' and r might be ' $(x)Fx \supset (x)Gx$ '. Is Oscar entitled to believe r ? Would a correct epistemic rule permit him (simply because he believes q) to believe r ? The answer, I think, is clearly no. But RGS 1 implies that there is such a rule. Since ' $(x)(y)(Fx \supset Gy)$ ' does logically imply ' $(x)Fx \supset (x)Gx$ ', RGS 1 implies there is a rule that permits anyone to move from believing the former to believing the latter.

If the unacceptability of such a rule isn't evident to the reader, we can argue for its unacceptability by reference to Principle I. Suppose Oscar's belief in proposition q is justified, i.e., results from transitions that accord with correct epistemic rules. Suppose further (for the purpose of a *reductio*) that all rules generated by RGS 1 are correct. Then if Oscar goes ahead and believes r in the next moment, he will be making a transition that conforms with a correct epistemic rule. His belief in r will also have a history of transitions that conform with correct rules, and hence, by Principle I, his belief in r will be *justified*. But suppose that Oscar doesn't even believe, or realize, that q logically implies r . He believes r by sheer hunch or guesswork, or just by wishful thinking. Intuitively, then, Oscar's belief in r is *unjustified*, even though the formation of this belief conforms with putatively correct epistemic rules. This shows that something is wrong with our previous assumptions. The plausible conclusion is that the rule generated by RGS 1 isn't correct. But if this rule isn't correct, then the schema that generates it, RGS 1, isn't a correct rule-generating schema. So RGS 1 isn't an instance of a correct, purely logical, rule-generating schema.

The way around this difficulty may seem simple: just require *belief* in the logical implication. In other words, replace RGS 1 with RGS 2, also a schema in which epistemic rules are to be derived from truths of logic alone.

(RGS 2) For any propositions q and r , if $q \models r$, then $P(Bq \ \& \ B(q \models r)) / Br$.

Before examining the adequacy of RGS 2, consider a different possible alteration of RGS 1. Since merely believing the implication, as opposed to believing it *justifiably*, may seem insufficient, the reader might wonder why we don't propose RGS 2' instead of RGS 2.

(RGS 2') For any propositions q and r , if $q \models r$, then $P(Bq \ \& \ J(q \models r)) / Br$.

Here '*J*' represents 'believes justifiably'. This amendment, however, is formally inadmissible. Our epistemic rules govern transitions between (pure) cognitive states, but justifiably believing is not a (pure) cognitive state; it is an epistemic status of a cognitive state. Furthermore, there is no need for the justification-condition to be built into the rule. Since according to Principle I, justifiedness can only be acquired via a history of approved transitions, a belief in *r* that conforms with RGS 2 could only acquire justifiedness if the belief in $q \models r$ were justified, i.e., had a history of conformity with correct rules. So RGS 2, conjoined with Principle I, is as strong as RGS 2', but doesn't share its formal inadmissibility.

Let us turn, then, to RGS 2. Is it a correct rule-generating schema? Unfortunately, we can still construct an Oscar-type case with the same sort of difficulty as the one we raised against RGS 1. Suppose Oscar believes *q* (justifiably) and believes $q \models r$ (justifiably), and suppose that he now comes to believe *r* as well. But the *cause* of his coming to believe *r* is not the aforementioned beliefs, but something quite extraneous. Perhaps he believes it, once again, by mere guesswork or wishful thinking. Or perhaps he believes it because a wholly untrustworthy guru has uttered *r*, and Oscar is irrationally inclined to believe whatever the guru says. Then Oscar's belief in *r* isn't justified, yet it will be justified if RGS 2 is correct. Hence, RGS 2 must not be correct.

A defender of RGS 2 might not be entirely persuaded. How is it possible, he might say, for someone to believe *q*, to believe $q \models r$, and yet *not* to believe *r* as a result of these beliefs? I reply: Perhaps Oscar's belief in $q \models r$ is stored in memory, and doesn't get retrieved or accessed; perhaps he learns it some time before he learns *q*, and never puts the two propositions together.

A persistent defender of RGS 2 might still not concede the point. He might, for example, insist that the intended sense of 'believe' in RGS 2 is the "occurrent" or "conscious" sense. In that sense of 'believe', he might say, it is *psychologically necessary* that if you believe *q* and believe $q \models r$, then you will also believe *r* because of the former beliefs; you will necessarily "see" the connections between *q* and $q \models r$, on the one hand, and *r*, on the other. Given that situation, the belief in *r* will be justified.

Suppose we grant this contention. Then it is clear that RGS 2 is a correct principle only because of certain *psychological facts*! It is only a

contingent feature of the human cognitive system that makes RGS 2 work. For only that contingent fact guarantees (if anything guarantees) that a person will “see” the connection between q and $q \models r$ and infer r .

Someone might reply that it isn’t a contingent psychological fact that this inference is made; rather, it follows “analytically” from the concept of logical implication, or from what it is to believe a logical implication. A cognizer couldn’t be *counted* as *believing* $q \models r$ unless he automatically makes this sort of inference.

I don’t find this approach convincing. Compare the present example with a case in which someone believes $q \models r$, believes $r \models s$, and believes not- s . Suppose these beliefs do not automatically cause him to believe not- q . Does it follow that he doesn’t really understand logical implication? That he shouldn’t really be counted as believing the indicated implication-statements? Not at all. He may understand implication perfectly, but simply not *notice* the indicated relationship. After all, we cannot plausibly decree that a person doesn’t understand logical implication unless he recognizes *all* the logical relationships, however complex, that follow from certain implications. Such a decree would be excessive. Can we say, nonetheless, that a person doesn’t understand logical implication unless he always recognizes instances of Modus Ponens, e.g., always infers r from q and $q \models r$? If there is plausibility in this requirement, it comes from the fact that the relationship between q , $q \models r$, and r is a very *simple*, *obvious*, and *transparent* one. But simplicity, obviousness, and transparency are *psychological* notions, not purely logical ones. A relation is simple if it is *easy* for a cognizer to apprehend it. But *ease* is relative to cognitive capacities or structures. As Christopher Cherniak (1981) points out, we can imagine beings with different psychological compositions from ours, who find “easy” the logical relations we find “difficult”, and vice-versa. So whatever plausibility there is in saying that a person *must* infer r from q and $q \models r$ (if he is to be counted as believing q and $q \models r$) derives in part from certain contingent psychological facts, and we are not really deriving epistemic rules from *logic alone*.

There is another direction the Oscar case might take the discussion, though it results in the same conclusion. In response to the (second) Oscar case, one might concede that RGS 2 needs to be replaced, but perhaps the addition of a further belief-clause in the antecedent will do the job. That is, perhaps RGS 3 is a satisfactory schema.

- (RGS 3) For any propositions q and r , if $q \models r$, then $P(Bq \& B(q \models r) \& B((q \& q \models r) \models r) / Br)$.

But why should anyone think this third belief-clause will solve the problem if the second one didn't? Won't another Oscar case appear, forcing still another belief-clause to be inserted into the antecedent? Doesn't a vicious regress threaten, of just the sort Lewis Carroll (1895) presented in "What the Tortoise Said to Achilles"? Where should the regress be stopped, and what is the rationale for stopping it at any particular place?

It isn't obvious to me *where* the regress should be stopped (indeed, perhaps *none* of the schemas RGS 1, RGS 2, RGS 3, etc., is correct). The only rationale I can envisage, however, for stopping at any of these junctures would depend in part on assumptions about human psychology, on the sorts of relationships it is *easy* for human cognizers to detect, or the sorts of relationships they regularly *do* detect. Thus, none of these schemas succeeds in showing that epistemic rules can be derived from *logic alone*, without reliance on *psychology*.

The whole issue of the relation between logic and epistemic rules may be put more perspicuously by focusing on the possibility of unconditional rules for believing truths of logic. Many confirmation theorists have said that any tautology (or valid sentence) should be assigned a probability of 1.0, suggesting that a cognizer is rationally justified in believing a tautology, no matter what else he believes. If this were right, it would apparently be an instance of epistemic rules following from logic alone. So let us explore these kinds of cases.

How can unconditional epistemic rules be represented in terms of our cognitive-state-transition framework? Let 'V' represent a *universal* cognitive state, i.e., a *partial* state that a cognizer is automatically or vacuously in, no matter what other cognitive states he is in at the time. Then an unconditional permission to believe a proposition q may be written as: " $P(V/Bq)$ ". This says that one is permitted to make a "transition" from any cognitive state whatever to a belief in q . The idea that a person is automatically entitled to believe a truth of logic can now be expressed as the following schema:

- (RGS 4) For any proposition q , if $\models q$, then $P(V / Bq)$.

Unfortunately, this rule-schema is wholly unacceptable. There is no plausibility in the idea that a belief is *justified*, as RGS 2 and Principle I

jointly imply, *simply* because the believed proposition is a logical truth. There are many epistemic situations one can be in vis-a-vis a proposition that makes it unsuitable for belief even if it is, in point of fact, a logical truth. Extreme cases are higher-order truths of logic that have never been proved by anyone (or shown in any other way to be true). Surely nobody is justified in believing these propositions. Or consider a novice who contemplates a well-established logical truth but who hasn't himself learned by authority that it is true, hasn't used any algorithm to establish its truth, and cannot see intuitively that it is true. Surely such a novice isn't justified in believing it. The effect of RGS 4 is to conflate the semantic modal status of a proposition, considered in itself, with the epistemic status a cognizer has in relation to that proposition. While such a conflation has been all too common in past epistemology, it must be resisted.

If necessary, the point can be illustrated with another Oscar case. Suppose Oscar believes a certain tautology because of his irrational trust in a guru who just happens to have uttered that tautology. Then Oscar's forming this belief would conform with a correct epistemic rule, and according to Principle I, the belief would be justified. Moral: RGS 4 isn't a correct schema.

Some replacement for RGS 4 *might* be acceptable. RGS 5, for example, is much more attractive.

(RGS 5) For any proposition q , if $\models q$ and it is *obvious* that $\models q$, then $P(V/Bq)$.

However, as remarked earlier, the notion of *obviousness* isn't a notion of logic; it is a *psychological* notion. What is obvious for human beings (or any other beings) depends on cognitive endowments. So schema RGS 5 doesn't support the idea that epistemic rules may be derived from logic alone. On the contrary, it illustrates how epistemic rules depend on facts of psychology.

5. EPISTEMIC RULES AND COGNITIVE OPERATIONS

A careful reader might be puzzled by my rejection of some of the foregoing rule-generating schemas but my (qualified) acceptance of Principle II. According to this Principle, what makes epistemic rules correct, or right, is that they promote true belief and error-avoidance

better than any alternative. So rules that promote belief in truths and only truths seem to be right, or correct. But that is precisely what holds of rules generated by some of the rejected schemas. Consider RGS 4, for example. Each rule it generates licenses belief in a truth and only a truth. Shouldn't it therefore be a correct rule, according to Principle II? Isn't there an inconsistency, then, in my endorsement of II, on the one hand, and my rejection of RGS 4?¹²

There is indeed an inconsistency, and it shows that some sort of revision of Principle II is in order. In considering a suitable revision, notice first that Principle II places no constraints on *how* epistemic rules should promote true belief and error-avoidance. One may have assumed, quite understandably, that such rules would specify *general strategies* of good cognition, and this is indeed what one expects normative epistemology to do. But Principle II does not explicitly say this. Since Principle II would sanction RGS 4, it allows epistemic rules to enumerate individual truths (here, logical truths) that may be believed. Similarly, as far as Principle II goes, the best set of epistemic rules for contingent propositions *could* turn out to be a very long list of instructions to believe an enumerated set of propositions. The epistemologist may take his entire "encyclopedia of knowledge" and simply prescribe or permit belief in all of those propositions. Clearly, this is not in the spirit of epistemic rule-giving. So Principle II does not do exactly the job we want it to do.

Why doesn't a set of instructions to believe individually enumerated propositions fit the spirit of normative epistemology? The aim of epistemology is to specify good procedures for getting information *on one's own*, not simply to *impart* information. The epistemologist qua epistemologist isn't an informant or purveyor of truths. He is interested in describing general *methods* for getting truths, including methods that can be applied to *new* problems or questions with which the epistemologist himself may not be concerned. To put the point another way, the epistemologist is interested in characterizing or specifying *intelligent* or *rational* ways of proceeding, and this is not accomplished simply by saying which propositions are truths (and therefore to be believed).

This discussion focuses on the need for *generality* in epistemic rules. But there is another important point that emerged from the problems besetting RGSSs, 1, 2, 3, and 4. Recall that for all (or many) of these schemata, "Oscar counterexamples" were constructed. Oscar-cases are cases in which a cognizer has beliefs that satisfy the ante-

cedent of one of the generated rules, and also forms a new belief that satisfies the consequent; but the *causal process* that *produces* the new belief doesn't (intuitively) confer justifiedness. Thus, while a 'transition' occurs that conforms with a putatively correct rule, the resulting belief isn't justified. What we learn from the Oscar cases is that the *justificational status* of a belief depends on *why* it is held, i.e., on what *causes* it (or causally sustains it) (Goldman, 1979). Our 'transition' rules say nothing whatever about *causes* of belief. They permit a belief to be formed if certain prior beliefs are held, independently of whether there is any causal relation between these succeeding beliefs, or what kind of causal relation there may be. The Oscar counterexamples suggest that such permission rules cannot succeed. What is needed are not rules that permit states at one moment to be *followed* by states at the next moment (which is all 'transition' rules do), but rules that specify *operations* by which certain cognitive states will *produce* other cognitive states. These operations, of course, must be *psychological* or *cognitive* operations.

If we amend our framework so that epistemic rules are construed as specifying cognitive operations, or combinations of operations, I believe this will handle both the problem of *causal processes* and the problem of *generality*. The operations are to be conceived of as operating on prior cognitive states and materials, and hence as producing, i.e., causing, new states from old. Among these outputs would be beliefs and other doxastic states. Furthermore, these operations will be quite general in the sense that they can perform transformations on all sorts of different contents. Presumably, (some of) the same psychological operations of thought are employed for thinking about and believing all sorts of propositions, both contingent propositions and noncontingent propositions, both truths of logic and falsehoods of logic.

What are examples of cognitive operations that might occur in epistemic rules? Here is a brief list of some possibilities (cf. Miller and Johnson-Laird, 1976; Posner, 1973). First, there is the directing of attention to selected mental representations. This includes attention to items in the perceptual field as well as items in working memory. Closely associated with the management of attention is the process of search. Either the perceptual field or memory can be searched for various sorts of things. Consequent upon memory search, there is typically retrieval from memory. Items retrieved may range from names and faces to plans, routines, algorithms, 'frames', and the like. In addition to memory search and retrieval, there is storage in memory.

This involves labelling or tagging the item to be stored, perhaps identifying it in terms of its relations to other mental representations. More generally, all sorts of operations may exist for establishing relationships between representations. One can try to determine a match or mismatch between items, including an attempt to determine whether a perceptual element has been previously encountered ('recognition'), or an attempt to find shared features between two perceptual elements. Manipulations of mental patterns and structures may occur, e.g., segmenting a pattern into parts, grouping parts into a whole, or establishing an order among elements. Finally, there may be operations of forming abstractions or prototypes from perceived instances, and of forming new combinations or scenarios from preexisting components.

How good a list this is doesn't matter for present purposes. The framework I am advocating has no commitment to any particular doctrines in the psychology of cognition. *Substantive* normative epistemology will have to be quite specific, of course, but at this point I am just trying to sketch the right *meta-theory*. However, additional concrete examples will be explained in the next section.

So far it sounds as if epistemic rules will take one of the following forms:

$$P(o_1 - o_{17} - o_{48}), \text{ or } O(o_{97} - o_{142}), \text{ or } N(o_{326} - o_{63} - o_8)$$

where '*P*' is the epistemic permission operator, '*O*' is the epistemic obligation operation, '*N*' is the epistemic prohibition operator, and the concatenated o_i 's represent various combinations or sequences of detailed mental operations. But the rules will probably have to be more complicated than this. For one thing, which combination of operations is appropriate or inappropriate depends on the kind of cognitive *task* at hand, and the *prior state* of the cognizer. For example, if the cognizer wishes to solve a certain problem or puzzle, how he should proceed depends on whether he thinks he has ever learned a relevant algorithm, or whether he has ever encountered similar problems in the past. If so, a search in memory for the relevant algorithm or past encounters will be appropriate; otherwise not. Second, it will not do simply to say *which* operations should be performed; rules will have to specify the antecedent states or contents *to which* certain operations should be applied. If we retain the natural idea that a total cognitive state at a given moment is composed of many items, elements, or sub-states, then

certain operations should be applied to certain of these items or sub-states and other operations should be applied to certain of these items or sub-states and other operations to other items or sub-states. Thus, epistemic rules will doubtless be very complicated.

It should be noted that some of the operations in question will be *voluntary* and others *involuntary* or *automatic*. Both kinds are of interest and may appear in epistemic rules. Focusing of attention, for example, is often voluntary; what is noticed or apprehended as a result of such attention is not. Searching memory for an item that fits a specified 'cue' may sometimes be voluntary; the response produced by memory is involuntary. Trying to 'match' one mental content with another may be voluntary; but the judgment of whether there is a match or mismatch is automatic. Combinations of operations can be evaluated as (epistemically) good or bad whether or not they are voluntary. For example, the operations that constitute 'wishful thinking' or 'hasty generalization' may well be involuntary. Still, we can classify such operations as epistemically bad; the belief-outcomes of these operations will be *unjustified*.

We are now ready to contemplate a revision of Principle II. However, in the end I do not think that this principle needs to be revised. All we need is a revision in our *conception of an epistemic rule*. Instead of construing an epistemic rule as one that governs cognitive-state *transitions*, we should conceptualize epistemic rules as governing cognitive *operations*, especially combinations of such operations. With this conception in mind, Principle II can stand as written. (It should be recalled, though, that I do not endorse this principle without qualification.) Unlike Principle II, however, Principle I does make reference to cognitive-state "transitions". So the wording of *this* principle does need to be changed. We may simply substitute Principle I':

Principle I': *A person's doxastic state D (at time t) is justified if and only if D results from a history of cognitive operations that conform (are permitted by) the correct epistemic rules.*

To illustrate the impact of this change, return to our discussion of unconditional permission to believe truths of logic. Such an unconditional permission rule would license belief in a logical truth no matter what mental process produces the belief. But this is inappropriate, even in the case of "obvious" truths of logic. If a person goes through *wrong* mental operations in coming to believe 'If *p* then *p*',

then his belief isn't justified, despite the obviousness of this truth. Now our new "cognitive operations" framework for rules would accommodate this point. No blanket permission to believe a specified proposition, or set of propositions, would be issued, since such a permission is silent about cognitive operations. The rules would now have to be rules that license suitably selected operations (i.e., ones that promote truth-acquisition and error-avoidance). If one uses appropriate operations, a belief in 'If p then p ', or other trivial truths of logic, would be easy to arrive at. But a person will be *justified* in such a belief only if he arrives at it *via* appropriate operations. No entitlement to such a belief can be granted without reference to, or restriction on, such operations.

It is common to portray epistemology and philosophy of science as dealing with 'rules', 'methods', 'procedures', or 'principles' for inquirers and scientists to use. Our framework retains this idea. But many things that might be called 'methods' or 'rules' of proper procedure in the sciences would not qualify as epistemic rules on my conception. Each separate science, for example, typically develops its own special set of techniques, methods, or algorithms for dealing with problems in its domain. Formal sciences develop algorithms such as procedures for deriving square roots and doing long division; truth-table and truth-tree techniques for checking validity; and rules for regression analysis or analysis of variance. Empirical sciences develop all sorts of laboratory techniques for manipulating their chemicals, their cells, their particles, and their instruments; they develop methods for interpreting the significance of experimental results. While such techniques or methods are valuable pieces of intellectual advice in their appropriate spheres, they do not count as 'epistemic' rules. My conception of epistemic rules – a conception that fits, I believe, the central tradition in epistemology – is one of *domain-independent* rules: rules of operation that are applicable across all sorts of different subject-matter, or "content". What gives unity to the rules, on this view, is their concern with psychological operations. However wide and various are the topics to which the human mind can address itself, its modes of addressing these topics employ the same basic set of psychological tools. Epistemic rules concern themselves with the optimal use of these tools.

Perhaps this conception of epistemic rules is a bit too extreme. If so, a more moderate approach is readily available, and compatible with the arguments I have been giving. Rather than say that *all* epistemic rules are concerned exclusively with cognitive operations, this thesis might

be restricted to (what we may call) *fundamental* epistemic rules. Other methodological rules may also be contained as epistemic rules, but not as fundamental ones. On this approach, our way of linking justification with epistemic rules should be reformulated in terms of fundamental epistemic rules. Our argument for the relation between epistemology and psychology, however, will still stand. As long as fundamental epistemic rules must have the character I have imputed to them, there will be an important role for psychology to play in normative epistemology.

6. WHAT COGNITIVE STATES AND OPERATIONS ARE AVAILABLE?

A brief resume of our argument is now in order. Epistemic rules, or fundamental epistemic rules, will be concerned with cognitive states and cognitive operations. The right set of rules will be that total set of rules such that (a) human beings can conform with that set, and (b) conforming with that set would maximize the attainment of epistemically valued ends. Given this conception of epistemic rules, a paramount question is: What cognitive states and operations are available to human beings? What combinations of such operations could human beings instantiate, or realize? These are the questions that should prompt epistemology to seek help from psychology. They are the questions that make psychology relevant.

Will epistemology become, on our view, a branch of the psychology of cognition? Not at all (not even *individual* epistemology). Even a full (and accurate) set of answers to the above questions would not determine the correct set of epistemic rules. Psychology cannot do this on its own. For one thing, a choice of a right-making characteristic must be made, and this falls outside the domain of psychology (though even here, as we shall see, psychology may well be relevant). Second, if this characteristic features truth and falsity, as Principle II suggests, some nonpsychological inquiries will be needed to help decide *which* of the available mental operations best promote these ends. Finally, even if there is a small and relatively well-defined set of elementary cognitive operations, the combinations and permutations of such operations will not be small, nor even clearly bounded. The candidate sets of operations will not "fall out" from a list of elementary operations. Ingenuity will be needed to design optimal combinations of operations, with an

eye to promoting truth and the avoidance of error. This task does not fall to psychology alone. A mix of disciplinary inputs will be required, both logical-philosophical and psychological.

Our focus, however, is on the inputs from psychology. How could psychological investigation of cognitive states and operations have a bearing on epistemic rules? I have already suggested a short list of possibly relevant operations, but now let me explore some other possibilities in a bit more detail, to put more flesh on our abstract skeleton.

One question of general importance is whether there is a single *medium* of cognition, a unitary “*language of thought*”, as Fodor (1975, 1978, 1980) and Pylyshyn (1980) contend, in which all cognitive operations take place. On the assumption that cognitive states and operations use *some* sort of symbols, it is important to decide what kind or kinds of symbols there are. If there were a unitary language of thought, all cognitive operations would operate on symbols in that single language, or medium (“mentalese”). Since the specification of operations may well have to be couched partly in terms of the symbols they manipulate or transform, a single mental language or code would simplify the description or formulation of such operations.

In opposition to the *single-code*, or single-medium, theory of cognition is the *multiple-code* theory. This approach is usually combined with the assumption that some of the representational codes are *modality-specific*, i.e., are linked to vision, audition, proprioception, etc. Such modality-specific codes are allegedly at work in *imagery*, and perhaps in (long-term) *memory* as well. Adherents of imagery need not, in my opinion, be saddled with the claim that these codes are *picture-like*, that they represent by means of *resemblance* (cf. Dennett, 1969, Chap. 7 and 1978). Furthermore, they can be countenanced even if one thinks that they cannot serve as a vehicle of thought. (This seems to be Fodor’s (1975, pp. 191–193) tactic for dealing with imagery.)

Without trying to *settle* these complex issues, consider what impact the existence of multiple codes, or imagery, might have on epistemic rules. If there are multiple codes, or if thought is accompanied by sense-related images, certain of these codes or images may be epistemically helpful on certain occasions. It may be beneficial to use certain codes when tackling certain intellectual problems and other codes when tackling other problems. Epistemic rules could be tailored to promote these strategies.

To see how imagery can affect success in solving a problem, consider an example from Lawrence Powers (1978). Suppose you are given this problem: "What is a four-letter word ending in the three letters 'e', 'n', 'y'? Many people try to solve it by going through the alphabet looking for the missing first letter. When the problem is presented *orally*, one pronounces 'eny', as ENNY and assumes that the four-letter word will rhyme with ENNY. One goes through the alphabet imagining "*in one's mind's ear*" how AY affixed to ENNY would sound, how BEE affixed to ENNY would sound, etc. One constructs the following auditory images: AY-ENNY, BENNY, SENNY, or KENNY, DENNY, EE-ENNY, FENNY, etc. No word is found that sounds like a word of English. So one fails to solve the problem. There is a correct solution, however, the word 'deny'. This answer is more likely to be found if *visual* imagery rather than *auditory* imagery is used to represent the possibilities. If one *visualizes* what an 'a' prefixed to 'eny' would look like, what a 'b' prefixed to 'eny' would look like, etc., one is unlikely to overlook the fact that 'd' prefixed to 'eny' is a legitimate English word.

Any attempt to incorporate the use of imagery into epistemic rules will doubtless be complicated, not only because different problems lend themselves to imagery in different ways (or not at all), but because of individual differences in imaginal powers. Still, both introspection and well-known experimental results obtained by Shepard, Kosslyn, and others indicate that imagery plays an important role in thought and should not be neglected by epistemology.

The issue discussed so far concerns the unity vs. multiplicity of the medium of thought. A second issue concerns the unity vs. multiplicity of the '*location*' of belief. In the dominant epistemology of the analytic tradition, the unity of the notion of belief is taken largely for granted. A distinction has been drawn in the philosophy of mind between *occurent* belief and *dispositional* belief, but little or no epistemological use has been made of this distinction. Philosophers draw this distinction in roughly the following way. There are many propositions I can be said to 'believe' that I am not currently thinking about – e.g., propositions about world history or my own personal past. I believe these propositions, however, because I am *disposed* to assent to them, if they should come to my attention. There are other propositions, however, that are currently active in my mind, that I am consciously accepting as true – e.g., that there is a typewriter before me, and that my fingers are depressing its keys. These propositions are *occurrently* believed.

This notion of different kinds (or senses) of belief is connected with a common theme in cognitive psychology, i.e., that information passes through different *stages*, or is 'held' at different *locations*, in the mind. Of course, if and when the term 'location' is used (and it is more my term than a term of psychologists), it is understood *functionally* rather than neuroanatomically. However, psychologists do not talk in the language of 'belief'; they express the theme I have in mind in terms of different *memory stores*. Three such stores are frequently postulated. First, there is 'iconic' storage, where sensory information is held very briefly, up to a second or two at most. Second, there is short-term memory (STM), which can hold information somewhere on the order of 30 seconds, but where information rapidly decays if it isn't 'recycled' by repetition or rehearsal (Klatzky, 1980). Third, there is long-term memory (LTM), where information may be stored almost indefinitely, and where decay is much less rapid. I believe that the philosopher's distinction between *occurrent* and *dispositional* belief corresponds *roughly* to the difference between STM and LTM. An *occurrent* belief is some state of STM; a *dispositional*, or "standing", belief is a state of LTM. (The correspondence isn't perfect because not everything stored in STM is conscious, or is present in focal awareness. But I shall downplay this point.) Thus, the same proposition can be believed, or 'held', in two different (functional) locations. "Recall" consists of getting a proposition (or other item) from LTM into STM (though the item isn't lost from LTM in the process).

Such a model has important implications for normative epistemology. Our normative epistemology has tentatively chosen true "*belief*" as an aim of cognition, as a valued end to be sought by epistemic rules. Which kind of belief, though, is intended? Is value only attached to conscious true belief, to truths held in STM? Or is value attached as well to having a truth in LTM, even if it is difficult to retrieve? We commonly speak of "possessing" the truth on a subject, but what kind of possession is this? Is it like having the truth at hand? Or having it on file, or in a safety deposit box? Or, as in Plato's metaphor, like having a bird in an aviary? What if a truth can only be activated from LTM with great effort? Does this affect the extent of its value?

Epistemologists have discussed the relative value of different propositions. Stronger, simpler, or more explanatory theories are often said to be more desirable. The question I am raising has nothing to do with the relative value or importance of different *propositions* but with the

relative value of the *mental lodging* of a given proposition. This problem has been neglected, because epistemologists have failed to recognize that different locations in the informational network of the mind confer different causal powers. If a belief is only lodged in LTM, it may have no direct impact on current decisions or inferences. For example, if I am trying to decide whether to go to the University Library this Sunday morning my 'knowledge' that the Library doesn't open until 1:00 on Sunday will not affect this decision unless this information is activated, retrieved, or recalled – i.e., brought into STM. Similarly, even if I have a stored belief in a significant theoretical truth, this won't help me to solve a problem it bears on unless I access the theory and apply it to the problem at hand.

If only activated beliefs are 'directly' of use in decision-making and problem-solving, why not enjoin cognizers to activate *all* their beliefs? The standard answer in cognitive psychology is that short-term memory (STM) is a *limited-capacity* domain, so only a small number of items ($7 + 2$, in the classical formulation) (Miller, 1959) can be held there at a given moment. A basic problem for cognitive operations is which items to maintain in consciousness during a problem-solving activity and which items can be 'released' without major risk. One dimension of intelligence may consist in wise choices of information-retention in problem-solving tasks. This topic should not be neglected by epistemic rules.

Because of the threat of informational overload in STM, the ability to group material into larger or more abstract units, or "chunks", is a valuable asset. The formula of $7 + 2$ is intended to apply to such chunks, so using larger chunks retains more information. There is reason to think that what distinguishes experts from novices in an intellectual domain is the use of larger, more abstract, or more relevant chunks. For example, the chess expert may be able to see a chess game in terms of more global groupings of pieces.¹³ And what distinguishes a sophisticated mathematician from a novice is the ability to grasp the overall strategy of a proof in a suitable mental structure. A novice reading a proof for the first time may get lost in the maze of lemmas and sub-proofs; when he has finished reading the proof he may not understand it as a whole, though he has followed each separate step. The expert mathematician, by contrast, not only follows each individual step, but intuits the larger structure. This mental difference makes him better justified than the novice in believing the theorem.¹⁴

Because analytic epistemology has focused so heavily on propositional content, it has tended to assume that as long as a certain 'corpus' of propositions is 'believed', no further attention to the mental embodiment is relevant. But since information-retrieval is essential, this is not correct. According to most current theories of memory, the retrievability of material from memory (LTM) depends on how that material is *organized*: what mental connections, links, or pathways there are between these items (cf. Anderson and Bower, 1973). Cognitive *organization* is a crucial element in epistemic success.

An illustration will help to clarify this point. Suppose that a group of American adults are asked about the 50 States of the Union. These people share the following characteristics: given any correct name of a State – 'Ohio', 'Nevada', 'Virginia', etc., – they would recognize it as correct, and given an incorrect state name – 'Instanbul', 'Ferdinand', 'The Pelopponesus' – they would recognize it as incorrect. Do they then share the same cognitive condition *vis-a-vis* this subject-matter? On the standard epistemological viewpoint, this would seem to be so. Since they would assent and dissent to the same suggestions, they appear to have the same set of beliefs about the States of the Union. Despite the similarity of their assent and dissent propensities, however, these people may differ sharply in their relevant cognitive condition. Suppose they are given the task of *listing* the 50 states (in three or four minutes). Some may have a systematic format for generating such a list, for example, the visual map of the Union from which the states may be "read off"; others may have a memorized alphabetical list. Still others may have no such algorithm at all, and would have great difficulty producing a complete enumeration. This latter group has no systematic way of connecting the various states, or state names, in their memory, and this *lack of organization* is crucially important to their performance on the specified intellectual task.

The centrality of organization can be illustrated not only in connection with retrieval, but also in connection with encoding and storing material in LTM. Epistemologists have often failed to appreciate that there is a problem of how to represent or encode a body of information. They often just make a blithe assumption that one somehow "grasps" relevant propositions. But complex information cannot just be "grasped"; it needs to be somehow *integrated* into one's prior mental structure, i.e., to be *organized* within antecedently existing memory networks. This process of establishing *connections* with preexisting

structures is also essential to, or at least facilitative of, LTM storage.

These points are made concrete by the following selection from the literature on story understanding, taken from an article by John Bransford and Nancy McCarrell (1975).

Bransford and McCarrell presented subjects with this passage, asking them to be prepared to (a) rate it for comprehensibility, and (b) answer questions designed to measure their recall for the passage.

The procedure is actually quite simple. First you arrange things into different groups. Of course one pile may be sufficient depending on how much there is to do. If you have to go somewhere else due to lack of facilities that is the next step, otherwise you are pretty well set. It is important not to overdo things. That is, it is better to do too few things at once than too many. In the short run this may not seem important but complications can easily arise. A mistake can be expensive as well. At first the whole procedure will seem complicated. Soon, however, it will become just another facet of life. It is difficult to foresee any end to the necessity for this task in the immediate future, but then one never can tell. After the procedure is completed one arranges the materials into different groups again. Then they can be put into their appropriate places. Eventually they will be used once more and the whole cycle will then have to be repeated. However, that is a part of life.

Subjects to whom this passage was presented "cold" rated it very low in comprehensibility, for it is difficult to *integrate* this passage into one's antecedent body of knowledge. These subjects also showed poor recall, presumably for the same reason. By contrast, some subjects were told before hearing the passage that it is about *washing clothes*, and this made a dramatic difference both in their comprehensibility ratings and on their recall tests. (Please re-read the passage now with the knowledge that its topic is washing clothes.) Without knowing the topic, the passage is hard to piece together. It seems to be a series of unrelated items, and no coherent picture emerges. Once one knows it is about washing clothes, however, each sentence forges links with prior information about laundering, and the passage is integrated with that preexisting knowledge. The feeling of understanding is greatly facilitated, and so is subsequent retrieval. In short, the passage "means more" as it is read, and this greater *meaning* enhances the storage process.¹⁵

Now once we appreciate the importance of organization for memory, and the need for new material to get incorporated into the old, the standard epistemological model of belief acquisition must be revised. No longer can we view information as encoded *atomistically* into the mind, proposition by proposition. Nor can it be assumed that memory is

retained in propositional units, even if its verbal *expression* is propositional, i.e., sentential. In short, our conception of what cognitive *states* people are in, and what *operations* get them into those states, may be substantially changed by cognitive psychology.¹⁶ In any case, whether the "atomistic", propositional model is ultimately correct or incorrect, psychology can provide the best guidance on the topic. Since normative epistemology has a large stake in the question of what cognitive states we can be in, and what cognitive operations will get us into those states, normative epistemology will want to appeal to the best theories on these subjects psychology can offer.

As a final example, let's look at the topic of change, or revision, of belief. Consider the studies by Lee Ross (1977, 1980, and forthcoming) and his colleagues on belief "perseverance". Ross and co-workers did a number of studies in which subjects were initially given totally false and deceptive information about their performance at certain tasks, or the performance of someone they observed. For example, either a subject or an observed peer was given the task of distinguishing authentic from fake suicide notes. The subject was falsely informed that the performer did very well, or very poorly. Later, the subject was "debriefed", i.e., told that the initial information had simply been concocted, and there was no reason to infer anything about the performer's true abilities from these statements. Nonetheless, despite this total discrediting of the initial statements, subjects persisted in having beliefs clearly influenced by the initial statements. This was revealed by answers to questions after the debriefing.

Ross and Lepper (1980) offer several possible explanations for this phenomenon. One explanation is that subjects seek and 'discover' causal explanations of the putative performance. For example, a subject who suddenly finds himself confronted with evidence of his superior or inferior ability at discriminating suicide notes might search for some aspect of his background or personality to account for such a talent or deficiency. The apparent failure might be explained, say, in terms of his own cheerful disposition as an impediment to the empathetic set the task demands. Once such an explanation has been 'found', later revelation that the original task outcome was contrived by the experimenters does not diminish the subject's confidence that he has a propensity to be deficient at this sort of task.

The search for causal explanations (relying on prior beliefs) fits the general picture of the human cognizer as trying to integrate new

information into his total cognitive scheme. Prior beliefs are 'confirmed' and strengthened by the new data, and additional inferences are drawn. This creates difficulties, though, when the data turn out to be spurious. The subject then has the hard task of reconstructing and correcting the many changes the data have wrought. This is the problem of "*backtracking*", which researchers in Artificial Intelligence have recently been exploring (Stallman and Sussman, 1977; Doyle, 1979). What operations need to be performed to keep track of past assumptions and revise conclusions affected by these assumptions when the assumptions prove false?

As presented here, the problem is mainly one of recall: can the cognizer *remember*, or keep track of, all the credences that past assumptions influenced? But I suspect there is a further problem. While we standardly speak of *dropping* or *abandoning* a belief – perhaps we have a '*blackboard model*' of the mind, in which sentences can easily be written or erased – it isn't clear that this is literally what happens. One of the going accounts of LTM depicts the process of acquiring belief as establishing associative *links* or *pathways* between relevant elements (e.g., concepts). Furthermore, it is questionable whether there is any way of *severing* or *disconnecting* these links. The links can only *decay*, or get *interfered with* by newly established links or pathways. Suppose, for example, that I used to think that a certain man with a funny chin is a mortician, but I am later apprised he is a banker. Initially, what I had was an associative link connecting (a concept, or image, of) the funny chin with (a representation of) being a mortician. When I am later told that this is incorrect, the chin-mortician link doesn't disappear. Rather, the chin-banker link is established, and perhaps a further link connecting the funny chin representation with the representation, 'my old information about this man was false'. Now, when I ask myself the question, 'What does the man with the funny chin do for a living?', the old link may no longer be excited (though it hasn't disappeared). It is just interfered with by the new association. Or perhaps all three links are triggered, and they jointly prompt me to say (to myself), 'He's a banker'. In either case, a change of belief has occurred. But the way it has occurred isn't by literal *deletion* of something (the old belief), but by *addition* of the new conflicting material.

How is this relevant to the perseverance studies? If the cognitive system (LTM, in particular) only works by superimposing new material, one cannot hope that subjects will literally expunge the old 'data'. What

they need is lively and forceful material to compete vigorously with the old. Interestingly, Ross and Lepper report that when *certain* debriefing processes are undertaken; specifically, when subjects were given insight into the psychological mechanisms that might create perseverance, and concrete illustrations of this, belief perseverance was *almost wholly eliminated!* The right kind of new material, in other words, has a cancelling effect, even if simple debriefing does not.

What is the moral of these reflections? We agree with Ross and colleagues, of course, that certain natural cognitive processes are 'counter-normative' in the sense that they promote error more often than one would wish. But what operations should be used to ameliorate the situation? The proper remedy depends, first, on precisely which natural operations are responsible for the problem, and second, on the kinds of operations available to combat it. Only psychological investigation can yield answers to these questions. If my hypothesis is correct that belief-change *really* consists (at a deep psychological level) in the addition of suitable conflicting material, rather than literal 'abandonment' of belief, then the best operations to ameliorate the problem may well be different from the operations that would otherwise be called for. It won't do simply to tell cognizers, "Forget what I told you earlier". Instead, some vivid and detailed story must be presented as a "countervailing force", or source of inference. This moral is applicable to change in scientific belief as well. *There*, however, the moral may already have been noted, in the observation (by Kuhn (1962) and others) that old scientific paradigms are only overthrown by new ones. The psychological underpinnings of this observation, though, and their widespread scope, may not yet be appreciated.¹⁷

NOTES

¹ These quotations from Piaget are presented in Bernard Kaplan (1971).

² The William James Lectures, presented at Harvard University in 1977, are thus far unpublished.

³ Cf. Carnap's notion of "explication", as presented, for example, in 1950.

⁴ This view of concepts contrasts, of course, with Frege's views, in which concepts are abstract entities, graspable by different individuals and hence mind-independent. The ontological status of such abstract concepts is problematic, however. Furthermore, even if we grant the existence of concepts in this sense, philosophical analysis aims at analyzing the concepts people *have*. The "having" of a concept is clearly a psychological affair.

⁵ Hilary Putnam, of course, has given interesting arguments against the attempt to locate meanings "in the head". See especially Putnam (1975).

⁶ Chisholm (1966 and 1976) states various epistemic principles, specifying when propositions are “beyond reasonable doubt”, “gratuitous”, “unacceptable”, etc. The normative status of these principles becomes apparent when Chisholm says, for example, that the category of unacceptability expresses epistemic “dispraise of condemnation” – to call a proposition unacceptable is to say, “Nay”, “Do not believe (it)”, (2nd edition, page 8). Bayesians formulate rules over time. For example, see Richard C. Jeffrey (1965).

⁷ The distinction between the theoretical and the decision-guiding function of rules is made by Holly M. Smith (A. K. A. Goldman, 1988). In two of my previous papers, Goldman (1978 and 1980), I discuss epistemic rules with the decision-guiding function in mind. Here I depart from that orientation.

⁸ The idea is often traced to F. C. Bartlett (1932).

⁹ The term “logic” is also used to refer to a subject called ‘*informal logic*’ often taught to introductory philosophy students. Unfortunately, there are no established truths of informal logic; indeed, it is quite unclear what the content of informal logic is, or should be. By contrast, formal logic has a well-defined content and set of truths. It is this body of truth I have in mind when I speak of ‘logic’ and when I ask whether epistemic rules can be derived from logic.

¹⁰ However, if psychology ultimately says that there are *no* states with propositional content – hence no belief states as usually construed – there is a problem about true-belief as the standard by which rules should be selected. In this contingency, different right-making characteristics may have to be chosen. Because of this possibility, psychology could be relevant to normative epistemology not only at the level of choosing rules, but also at the level of choosing right-making characteristics. We shall encounter another example of this in section 6.

¹¹ The issue here is analogous to one that arises in social theory. For example, John Rawls (1971) distinguishes between principles of “strict compliance” theory and principles of “partial compliance” theory, i.e., principles of justice designed on the assumption that they will be fully complied with and principles not based on this assumption. Similarly, Robert Nozick (1974) suggests that a theory of justice needs principles of “rectification” in addition to principles of acquisition and transfer of holdings. Principles of rectification are needed to handle cases in which the other principles have been violated. The question for epistemology is whether all epistemic rules should be based on the assumption that they will be followed, or whether all or some epistemic rules should make no such assumption, or indeed should provide specifically for past violations of rules.

¹² There is a tricky point of detail here. Are the rules generated by RGS 4 really the best among their competitors? Wouldn’t rules that *obligate*, rather than merely *permit*, beliefs in logical truths promote *more* true belief? One can conform with permission rules by abstaining from belief, yet this would result in less true belief than conformity with obligation rules to believe logical truths. This suggests that a correct set of rules would be the prescription variants of the rules generated by RSG 4. However, people aren’t capable of conforming with *all* such rules, at least on the plausible assumption that people cannot have infinitely many beliefs. Fortunately, the issue between prescription vs. permission rules of the sort in question can be ignored in this context. Both kinds of rules are objectionable, because they both sanction belief in logical truths simply because they are logical truths. Since Principle II would support *either* permission or prescription rules of this sort, this principle needs to be revised.

¹³ The work of Herbert Simon, W. G. Chase, and K. Gilmartin on this topic is reviewed by Klatzky (1980, pp. 311ff).

¹⁴ The notion of *degrees* of justification has not been explicated here. But it is an attractive notion that deserves a place in a more fully developed theory. Presumably, a better justified belief is one that results from a closer-to-optimal sequence of cognitive operations than a less justified belief.

¹⁵ This suggestion is consonant with the "depth of processing" theory presented in F. I. M. Craik and R. S. Lockhart (1972).

¹⁶ Challenges to the propositional or sentential approach are posed by Patricia Smith Churchland (1980) and by Paul M. Churchland (1981).

¹⁷ Versions of this 1981 paper were presented at Rice University, Northwestern University, University of Wisconsin-Madison, University of Wisconsin-Parkside, and, of course, at the University of Pittsburgh Workshop. It has changed, hopefully for the better, as a result of discussions at these institutions. I am also indebted to members of my graduate seminar at the University of Illinois at Chicago Circle, to Charles Chastain, and especially to Holly M. Smith for helpful comments.

REFERENCES

- Anderson, C. A., M. R. Lepper, and L. Ross: 1980, 'The Perseverance of Social Theories: The Role of Explanation in the Persistence of Discredited Information', *Journal of Personality and Social Psychology*.
- Anderson, John R. and Gordon H. Bower: 1973, *Human Associative Memory*, V. H. Winston and Sons, Washington, D.C.
- Bartlett, F. C.: 1932, *Remembering: A Study in Experimental and Social Psychology*, Cambridge University Press, Cambridge.
- Bransford, John and Nancy McCarrell: 1975, 'A Sketch of a Cognitive Approach to Comprehension', in W. B. Weimor and D. S. Palermo (eds.), *Cognition and the Symbolic Processes*, Lawrence Erlbaum, Hillsdale, N.J.
- Campbell, Donald: 1974, 'Evolutionary Epistemology', in Paul A. Schilpp, (ed.), *The Philosophy of Karl Popper*, Vol. 1, Open Court Publishing Co., LaSalle, IL.
- Carnap, Rudolph: 1950, *Logical Foundations of Probability*, University of Chicago Press, Chicago.
- Carroll, Lewis: 1895, 'What the Tortoise Said to Achilles', *Mind*, N.S. 4, 278-80.
- Cherniak, Christopher: 1981, 'Feasible Inferences', *Philosophy of Science* 48, 248-268.
- Churchland, Patricia Smith: 1980, 'Language, Thought, and Information', *Noûs* 14, 147-70.
- Churchland, Paul M.: 1981, 'Eliminative Materialism and Propositional Attitudes', *The Journal of Philosophy* 78, 67-90.
- Craik, F. I. M. and R. S. Lockhart: 1972, 'Levels of Processing: A Framework for Memory Research', *Journal of Verbal Learning and Verbal Behavior* 1, 671-84.
- Dennett, Daniel: 1969, *Content and Consciousness*, Humanities Press, New York.
- Dennett, Daniel: 1978, 'Two Approaches to Mental Images', *Brainstorm*, Bradford Books, Montgomery, Vt.
- Doyle, Jon: 1979, 'A Truth Maintenance System', *Artificial Intelligence* 12, 231-72.

- Fodor, Jerry A.: 1975, *The Language of Thought*, Thomas Y. Crowell Company, New York.
- Fodor, Jerry A.: 1978, 'Propositional Attitudes', *The Monist* **61**, 501-23.
- Fodor, Jerry A.: 1980, 'Methodology Solipsism Considered as a Research Strategy in Cognitive Psychology', *The Behavioral and Brain Sciences* **3**, 63-73.
- Goldman, Alvin: 1976, 'Discrimination and Perceptual Knowledge', *The Journal of Philosophy* **73**, 771-91.
- Goldman, Alvin: 1978, 'Epistemics: The Regulative Theory of Cognition', *The Journal of Philosophy* **75**, 509-23.
- Goldman, Alvin: 1979, 'What is Justified Belief?', in George S. Pappas (ed.), *Justification and Knowledge*, D. Reidel, Dordrecht.
- Goldman, Alvin: 1980, 'The Internalist Conception of Justification', in P. French, T. Uehling, and H. Wettstein (eds.), *Midwest Studies in Philosophy, Volume V, Studies in Epistemology*, University of Minnesota Press, Minneapolis.
- (Goldman) Smith Holly M.: 1988 'Making Moral Decisions', *Nous* **21**, 89-108.
- Goldstein, Ida and Seymour Papert: 1977, 'Artificial Intelligence, Language and the Study of Knowledge', *Cognitive Science* **1**, 84-123.
- Hamlyn, D. W.: 1967, 'Epistemology, History of', in Paul Edwards (ed.), *The Encyclopedia of Philosophy*, Vol. 3, Macmillan, New York.
- Jeffrey, Richard C.: 1965, *The Logic of Decision*, McGraw-Hill Inc., New York.
- Kaplan, Bernard: 1971, 'Genetic Psychology, Genetic Epistemology, and Theory of Knowledge', in Theodore Mischel (ed.), *Cognitive Development and Epistemology*, Academic Press, New York.
- Klatzky, Roberta: 1980, *Human Memory*, 2nd Edition, W. H. Freeman and Company, San Francisco.
- Kuhn, Thomas S.: 1962, *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago.
- Loftus, Elizabeth: 1979, *Eyewitness Testimony*, Harvard University Press, Cambridge.
- Miller, George A.: 1959, 'The Magical Number Seven, Plus or Minus Two: Some Limits in our Capacity for Processing Information', *Psychological Review* **63**, 81-97.
- Miller, George A. and Philip N. Johnson-Laird: 1976, *Language and Perception*, Harvard University Press, Cambridge.
- Moravcsik, J. M. E.: 1981, 'How Do Words Get Their Meaning?', *The Journal of Philosophy* **88**, 5-24.
- Nisbett, Richard and Lee Ross: 1980, *Human Inference: Strategies and Shortcomings of Social Judgment*, Prentice Hall Inc. Englewood Cliffs, N.J.
- Nozick, Robert: 1974, *Anarchy, State and Utopia*, Basic Books, New York.
- Pylyshyn, Zenon: 1980, 'Computation and Cognition', *The Behavioral and Brain Sciences* **3**, 111-32.
- Posner, Michael: 1973, *Cognition: An Introduction*, Scott Foresman, Glenview, IL.
- Powers, Laurence: 1978, 'Knowledge by Deduction', *The Philosophical Review* **87**, 337-71.
- Putnam, Hilary: 1975, 'The Meaning of "Meaning"', in *Mind, Language, and Reality, Philosophical Papers*, Volume 2, Cambridge University Press, New York.
- Quine, W. V. O.: 1969, 'Epistemology Naturalized' and 'Natural Kinds', in *Ontological Relativity and Other Essays*, Columbia University Press, New York.

- Quine, W. V. O.: 1974, *The Roots of Reference*, Open Court Publishing Company, LaSalle, IL.
- Quine, W. V. O.: 1975, 'The Nature of Natural Knowledge', in Samuel Guttenplan (ed.), *Mind and Language*, Oxford University Press, London.
- Rawls, John: 1971, *A Theory of Justice*, Harvard University Press, Cambridge.
- Ross, L.: 1977, 'The Intuitive Psychologist and His Shortcomings', in L. Berkowitz (ed.), *Advances in Experimental Social Psychology*, Academic Press, New York.
- Ross, L.: forthcoming, 'The Perseverance of Beliefs: Empirical and Normative Considerations', in R. A. Shweder and D. Fiske (eds.), *New Directions for Methodology of Behavioral Science: Fallible Judgment in Behavioral Research*, Jossey-Bass, San Francisco.
- Stallman, R. M. and G. J. Sussman: 1977, 'Forward Reasoning and Dependency-Directed Backtracking in a System for Computer-Aided Circuit Analysis', *Artificial Intelligence* 9, 135-19.

University of Arizona
Tucson, AZ 85721
U.S.A.

PART VI

THE MENTAL AND THE PHYSICAL

TWO VERSIONS OF THE IDENTITY THEORY

At present, functionalism and central state materialism are the two most popular and fully developed versions of the identity theory. I take any theory of mind to be a version of the identity theory if it includes the claim that each of the states referred to by mental terms could be described in physical terms, and is therefore a physical state. Although it has achieved the status of orthodoxy, the identity theory has been criticized repeatedly because of one class of phenomena. The problem concerns the qualitative character of sensation and perception. It is alleged that the presence of qualia in sensation and perception shows that functionalist and materialist analyses of mental states are simply wrong. In this paper, I hope to establish four claims. (1) Qualitative character is, at most, a problem for functionalism and central state materialism. It is not a problem for the identity theory in general. (2) Recent work in the philosophy of language, specifically, the contributions of Kripke, Putnam and Donnellan, enable us to formulate a version of the identity theory which successfully avoids the problem of qualia. (3) Indeed, current theories of reference enable us to formulate two versions of the identity theory, both of which avoid objections based on qualia. (4) The two versions of the identity theory I propose have three significant virtues. (a) They retain the advantages of functionalism and central state materialism. (b) They enable us to say about qualitative character what we need to say about it, no more and no less. (c) The theories are based on a more adequate theory of language than previous versions of the identity theory.

I. FUNCTIONALISM, CENTRAL STATE MATERIALISM AND QUALIA

When first thinking about the identity theory, one confronts a puzzle. If mental states are physical states, then when we discuss mental phenomena, we are actually talking about particular physical states. But how can it be reasonable to claim that we have been talking about specific physical states for centuries when, even now, we do not know which physical states various mental states will turn out to be? The identity theory has achieved

dominance as a result of the discovery of two plausible solutions to this puzzle. One has been proposed by central state materialists, the other by functionalists.

Central state materialists claim that, although the *designata* of mental terms are physical states, the “concept” of a mental state is not the concept of a physical state. Rather, it is the concept of a state which *typically* plays a certain causal role. We may analyze any mental concept, M , by correctly filling in the blanks in the following frame: M is the state which is apt to be produced by —, — · · · and/or apt for the production of —, — · · ·.¹ The solution to the puzzle of the last paragraph is that since mental terms apply to states solely in virtue of the states’ typical causes and effects, we may rightly claim that when using mental terminology, we are (unknowingly) talking about the physical states that, in fact, typically play certain causal roles.

By contrast, functionalists believe that mental terms refer to states through the functional roles they *actually* play. We call a state “pain,” for example, because it follows bodily injury, precedes crying and wincing, induces a desire that it should cease and so forth. Since mental terms refer to states solely through the states’ actual functional connections, it is reasonable to say that, when using mental terminology, we are referring to the physical states which have those connections.

This difference between central state materialism and functionalism can be seen more clearly by considering two examples. Suppose Ernie is in a state classified by physiology as an “ABC state.” Let’s assume that ABC states *typically* follow bodily injury, precede crying and groaning, induce a desire that they cease and so forth. For some reason, however, the usual causal connections are absent in this instance. Bert is in the opposite situation. He is in a state classified by physiology as “DEF,” and DEF-states (almost) never occupy the functional role of pain. In this case, however, Bert’s DEF-state has, by some fluke, the typical causes and effects of pain. A central state materialist would claim that Ernie is in pain, while Bert is not. Functionalists would say exactly the opposite.²

For the next section and the remainder of this one, I will ignore the differences between functionalism and central state materialism. Besides being varieties of the identity theory, the theories are in general agreement that mental terms refer to states in virtue of functional interrelations: the relations of mental states to inputs, outputs, and one another.

To support their claims about how mental terms work, materialists and functionalists often provide examples of “functional” or “causal” “analyses” associated with various mental terms. The analyses purport to list the causal connections through which the mental term refers to states. A deeply felt worry about functionalism, central state materialism, and so the identity theory itself, is that the analyses which have been proposed, or suggested, for sensation terms and perception terms seem inadequate, because they omit any reference to the qualitative character of these states. While no one seems to have a good general definition of qualitative character, it is easy enough to point to examples, and this is all that is required to state the objection. Using the typical case of “pain,” the objection is that functionalism and central state materialism are wrong, because we would *not* call something a pain even if it typically (or actually) follows bodily injury, precedes crying and wincing, and so forth, if it did not hurt, have a painful quality. Following Ned Block and Jerry Fodor, I will call this objection the “absent qualia argument.”³

How is an identity theorist to handle this argument? One way is to try to enlarge functional analyses of sensation and perception terms, so that they are invulnerable to absent qualia counter-examples. Sydney Shoemaker adopts this approach in “Functionalism and Qualia.”⁴ Shoemaker’s illuminating remarks about functionalism deserve a more extended discussion than I can offer here. Ignoring many interesting suggestions, I will consider only what I take to be the motivating principle of his strategy. Shoemaker confronts the aficionado of qualia with a dilemma: either the qualia involved in sentience are idle, or they perform some function in perception and sensation. If they are idle, totally lacking in causal connections, then it is not clear that we can know anything about them. At least, this would be true if we assume a causal theory of knowledge. Why would we think such qualities even exist, let alone play a criterial role in individuating mental states? If qualitative character is not idle, however, then

... we could give at least a partial functional characterization of the having of qualitative character by saying that it tends to give rise, in such and such circumstances, to those physical effects, and could not allow that a state lacking qualitative character could be functionally identical to a state having it.⁵

Shoemaker’s dilemma fails to catch the absent qualia argument, because it does not defend where the argument attacks. No one need claim that

qualia lack causal connections. Rather, the objection is that since states without qualia (seemingly) could perform the functions carried out by qualitative states, then even an enlarged functional definition will be vulnerable to counter-example. Shoemaker maintains that unless qualitative character is to be idle and unknowable, it must play some functional role. Call the roles played by qualitative character “Q-functions.” In effect, Shoemaker argues that all states having qualia must perform Q-functions. To meet the absent qualia objection, however, he needs to show that only states having qualitative character can carry out Q-functions.

If, presently, we can neither show how to enlarge functional or causal analyses so that they avoid absent qualia counter-examples, nor demonstrate that such an expansion must be possible, the obvious alternative is to try to meet this argument by pinning our hopes on the future. This strategy would concede that currently available causal analyses fail to provide adequate specifications of mental states, but it would retain the expectation that presently unknown future analyses will be able to block the counter-examples. While I will ultimately adopt something like this strategy, it appears to have an unacceptably high price. This strategy concedes that absent qualia counter-examples show that we do not refer to unknown physical states solely through their known causal inter-relations. Thus, in so far as functionalism and central state materialism are committed to the view that this is how we use mental terms to refer to unknown physical states, this strategy abandons these two versions of the identity theory. Further, this concession seems to leave the identity theory in the same position it was in thirty years ago, in need of some account of why it is reasonable to believe that, in using mental terms, we are actually referring to physical states. In the next section, I will show how recent work in the theory of reference provides a straightforward solution to the latter problem. In the final section, I will argue that, despite appearances, this strategy does not have the unwelcome consequence that all the philosophical labor that has gone into producing functional accounts of mental states has been in vain.

II. THE ABSENT QUALIA ARGUMENT AND THE NEW THEORY OF REFERENCE

An advocate of the identity theory must be able to answer two questions:

why is it reasonable to think that, in using mental terminology, we refer to physical states?⁶ what explains our ability to use mental terms consistently, given our present ignorance of physiology? As noted above, functionalism and central state materialism are important formulations of the identity theory at least partly because they seem to provide answers to these questions. Functional analyses, for example, seem to provide an identifying description through which we would be able to refer to unknown physical states. The absent qualia argument has seemed particularly devastating, because it has been widely assumed that an account of reference must present the (or an) identifying description through which we (tacitly) pick out the object of our remarks, and no description seems available for mental states other than a functional description. Hence the absent qualia argument can appear to show that reference to mental states is unique and mysterious: we refer to our own sensations and perceptions through some mysterious form of awareness of ineffable qualities.

Recent work in philosophy of language, specifically the causal theory of reference, suggests that a presupposition of this line of reasoning is false. As even its proponents admit, the causal theory of reference needs filling out. I will not try to work out any of the needed details, nor will I offer any defense of the basic theory. I wish to borrow two central claims from this large program, both of which I take to be adequately defended by others.⁷ The first, negative, claim is that it is possible for a person or a group of people to refer to at least some types of objects, even if no member of the group possesses an identifying description of the type. The second, positive, claim is that a group uses “*X*” to refer to *X*’s, even if they lack an identifying description of *X*’s, so long as the following types of systematic connections hold between their utterances of “*X*” and the presence of *X*’s: the members of the group use the word “*X*” to describe their surroundings when in the presence of *X*’s, when they wish to acquaint a neophyte with the word “*X*”, they show him *X*’s or describe the superficial properties of *X*’s to him, and so forth.⁸ The application of this plausible, if sketchy, account of reference to the identity theory is straightforward. An identity theorist would claim that even though we use mental terminology without being able to produce identifying descriptions of mental states, future physiology will provide a precise specification of pains, for example, something like “an organism is in pain if its *C*, *D*, *E*, *F*, or *G* fibers are firing at greater than *N* times per second.” Mental states will turn out to be brain

states, because it will turn out that our utterances of various mental terms were systematically connected with the presence of such physiological states.

Thus, even if he abandons functionalism and central state materialism, an identity theorist can appeal to the new theory of reference to explain how we might use mental terms to refer to unknown physical states. He can also appeal to recent work in philosophy of language to answer the second crucial question of how we use mental terms so consistently, given our ignorance of the physical states. Hilary Putnam's reflections on "conveying the meanings of terms" suggest two possible answers to the question of linguistic consistency. People may use a term consistently, nearly always correlating the term with a certain kind of object, merely through having acquired an ability to discriminate things of that type. Most adults are able to co-ordinate their utterances of "gin" with the presence of gin. Yet relatively few people know what liquids are gin, in the sense of being able to specify the properties a liquid must have to be classified as gin. Even if they cannot simply recognize objects of a particular type, speakers could still consistently refer to the objects if, when given the word, they are also provided with a short description of the objects to which the word is supposed to apply. The "stereotypes," as Putnam calls these brief descriptions, do not constitute identifying descriptions of the objects to which the word applies. Rather, a stereotype consists of only a few criteria for differentiating the objects from most of the rest of the speaker's environment.⁹

Once again, an identity theorist can simply apply this theory to the case of mental terms. We use mental terms to describe our own states and the states of others. When we use sensation terms to describe our own condition, the best explanation for consistent usage is that we have the ability to distinguish our own sensations. This fact has sometimes been expressed by saying that self-ascription of sensations is "criterionless." Although it has raised legendary problems for epistemology and previous philosophies of language, the ascription of mental states to others offers no particular challenge to the present theory. Just as the question of whether different chunks of metal were all really gold had to be resolved by chemistry, the perennial query about whether other people's pains are really the same as mine must be left to a future physiology. Meanwhile, we can apply mental terms to each others' states through the use of stereotypes; or, we may

simply discern the pains, emotions, and thoughts of others. More likely we attribute mental states to others in both ways.

Thus far, my argument has been that an identity theorist can defuse the absent qualia argument by conceding that functionalism and central state materialism do not explain how mental terms refer, because he can appeal to the causal theory of reference in their stead. However, philosophers who are perplexed about qualia may feel that this way of avoiding the argument merely postpones the problem. Although the difficulty about providing identifying descriptions has been circumvented for the present, it may be felt that the problem now shifts to the claim that future physiology will provide precise specifications of mental states. The idea would be that the existence of qualia somehow renders this claim false or dubious. As presented in Section I, the absent qualia argument confronted identity theorists with a clearcut objection. Their preferred accounts of how we use mental terms to refer to physical states were both vulnerable to persuasive counter-examples. Given only nebulous worries about whether physiology is up to the task of accounting for qualia, I do not think there is a great deal that an identity theorist can or must do to defend the theory.¹⁰ However, there is a well-known and non-nebulous objection to the identity theory based on qualia, that offered by Saul Kripke.¹¹ Having used Kripke's work to defuse the absent qualia argument, which is my main topic, I should at least indicate how identity theorists might handle this further objection based on qualia.

Because of lack of space, I will assume familiarity with Kripke's views about necessity, contingency, apriority, aposteriority, rigid and non-rigid designators. In brief, this is Kripke's argument. Like the scientific identity statements on which they have been modeled, the claims of the identity theory are necessarily true, if true at all. Thus, if "pains are the firing of C-fibers" is true, it is a necessary truth. The problem is that claims like the preceding seem contingent. Standard scientific identity statements also seem contingent, but their aura of contingency can be explained. The explanation is that we confuse the scientific identity claim with a related claim, which is contingent. For example, we conflate "heat is the motion of molecules" with "that which causes the subjective sensation of heat in me is the motion of molecules." The latter claim is contingent, for under different circumstances, some other phenomenon might affect us the way heat does. Hence, our intuition that "heat is the motion of molecules" is

contingent is compatible with the claim that this statement is necessary, because the intuition can be accounted for. According to Kripke, the critical difference between the model scientific identities and the claims of the identity theory is that our intuitions about the contingency of the latter cannot be explained away. It is naive to suggest that we confuse the necessary truth, "pain is the firing of *C*-fibers," with a contingent relative, "that which seems like pain to me is the firing of *C*-fibers," because "that which seems like pain to me is pain" is itself a necessary truth.

I think the feeling that the claims of the identity theory are contingent can be explained. I suggest the aura of contingency results from confusing epistemic possibility with metaphysical possibility, a confusion Kripke himself has done much to eliminate. At an unsophisticated level, the problem with a claim like "pain is the firing of *C*-fibers" is that presently we do not know whether it is true. Hence when presented with a narrative in which it is alleged that *C*-fiber stimulation exists in the absence of pain, we have no way of knowing whether that particular situation is possible. If pain is the firing of *C*-fibers, the situation would be impossible. Since we do not know this, however, we allow that the situation may be possible. In this case, we fall into the error of believing that "pain is the firing of *C*-fibers" is contingent, because we concede that it is possible to have *C*-fiber stimulation in the absence of pain, when our actual position is that, for all we know, *C*-fiber stimulation could exist without pain.

In Kripke's thought experiment, the same confusion arises, only on a different level. *Ex hypothesi*, pain is taken to be the firing of *C*-fibers, so we cannot make exactly the mistake just described. In this case, what gives rise to a sense of contingency is the reflection that the particular phenomenal quality associated with pain might be absent even though *C*-fiber stimulation is present. Of course, if pains are the firing of *C*-fibers and if pains necessarily have this particular phenomenal quality, then what we are imagining to occur would be impossible. Again, I suggest the problem is that we concede that it is possible to have *C*-fiber stimulation in the absence of the phenomenal quality, when our real position is that, for all we now know, *C*-fiber stimulation could exist without the phenomenal quality of pain.¹²

To sum up, I think that an identity theorist can meet the objection that, in drawing on the new theory of reference to block the absent qualia argument, he merely postpones the problem of qualia, either by shifting

the burden of proof or by specific replies to concrete difficulties. Unfortunately, supporters of the identity theory seem to have reasons of their own for being dissatisfied with this defense of their theory. What is to become of the functional analyses for mental states that have been offered during the last twenty or thirty years? The purpose of the next section is to show how this defense against the absent qualia argument still allows functional analyses to play a very important role in the philosophy of mind. My approach to this issue will be somewhat indirect.

III. TWO VERSIONS OF THE IDENTITY THEORY

Identity theorists may be divided into two groups. Many philosophers and psychologists believe that human behavior will never be explained in neurophysiological terms, either by laymen or experts. For the kind of information needed in dealing with human subjects, the neural level is thought to be too specific. Philosophers and psychologists of this persuasion would wish to claim that, in using mental terms, we are referring to states which are best understood as psychological states. It is not that these states could not be described physically for, in principle, it must be possible to specify the condition of every one of my cells at every moment in time. Rather, the claim is that a psychological description is more perspicuous than any physical characterization. I will call this version of the identity theory the "psychological theory."

If an identity theorist denies that psychology provides a useful taxonomy of the states we refer to in using mental terms, then he must advocate the view that these states are usefully classified by physiology. He need not defend the "type-type" position, but he must believe that physiological classifications provide a genuine taxonomy of mental states. One kind of mental state cannot turn out to be too many kinds of physiological states. Were this to happen, the correlations between our mental utterances and the physiological states would be too loose to justify the assertion that we have been talking about the physiological states. And if we are not talking about psychological states either, then why think that we refer to anything when using mental terminology? Hence, an identity theorist who denies the usefulness of psychological classifications is committed to the fruitfulness of physiological classifications of mental states. I will call this version of the identity theory the "physiological theory."¹³

The defense of the identity theory offered in the previous section was actually a defense of the positive claim of the physiological theory. Thanks to the new theory of reference it is possible to see how it could be reasonable to claim that mental terms refer to presently unknown physiological states. My concern is not to adjudicate between the physiological and the psychological versions of the identity theory. Rather, I want to argue that both versions can be defended against the absent qualia argument. Like central state materialism and functionalism, the psychological theory appears to be in danger from this argument. The theory maintains that, in using mental terms, we refer to states that can be characterized by psychology. To my knowledge, no current psychological theory provides any account of qualitative character. Hence, any presently available psychological description of sensations will be vulnerable to absent qualia counter-examples: (seemingly) it will be possible for states to meet the description, even though they lack qualia.

I think the solution is once again to sharpen the psychological theory into the claim that, in using mental terms, we refer to states that can be characterized by a *future* psychology. A psychological theorist should admit that, presently, we are unable to provide identifying descriptions of mental states, in either psychological or physical terms. He can explain the consistency in our use of mental terminology exactly as we did in Section II. To defend the psychological theory, what remains to be shown is how we can use a mental term to refer to a class of physical states which do not belong to a physiological kind, but to a psychological kind. That is, a psychological theorist must argue that a future psychology can play the role assigned to physiology in Section II. At this point, we need to be clearer about what is meant by "psychology," since physiological psychology is also a kind of psychology. When philosophers and psychologists reflect on the inevitability of psychology, I believe they intend "psychology" to refer to a recognizable extension of our everyday psychological views and the experimental results which have already extended our common views. They believe, for example, that human actions will always be explained in terms of beliefs and desires, and that beliefs and desires will always be explained partly by reference to their objects. I will call our everyday psychological views, plus the experimental results which have extended those views, "traditional psychology." "Psychological state" will be used to indicate a state as classified by traditional psychology. A re-

cognizable extension of traditional psychology would include the following: most of our present beliefs about the functional relations among mental states, stimuli and behavior; additional claims of the same type; some obvious extensions of current ideas; and, perhaps, some novel additions as well. The psychological theory would then be the position that the states we refer to in using mental terminology can be characterized precisely by an extended traditional psychology.

An extended traditional psychology could not play exactly the role assigned to physiology in Section II. The replacement of non-identifying, superficial descriptions by physiological descriptions would involve a shift to the micro-level. In changing from presently available descriptions to descriptions provided by an extended traditional psychology, we would not abandon the present level of analysis. When presenting his account of reference, Putnam is obviously thinking of micro-level theories as furnishing the exact descriptions of the phenomena. I believe he emphasizes these types of theories, because the causal theory of reference is most obviously correct for natural kinds, and natural kinds are widely believed always to share micro-structural properties. However, I see no reason why this should be the preferred level of description for all phenomena. If an extended traditional psychology can be a true theory, that is, a coherent set of laws which describes, explains and defines a range of phenomena,¹⁴ then it could provide a precise characterization of psychological states. Our mental utterances could be systematically correlated with the presence of these states. Hence mental states could turn out to be psychological states.

Once again, an aficionado of qualia is going to question whether this way of avoiding the absent qualia argument simply postpones the problem. Proponents of the absent qualia argument will be quick to point out that they believe their argument will be effective against *any* functional specification of sentient states, not just those which have been proposed thus far. I think psychological theorists can make at least two different responses to this objection. One reply is to deny the assumption that an extended traditional psychology must specify mental states solely by reference to functional inter-relations. A recognizable extension of traditional psychology could allow a minor role to physiological or other structural descriptions. The other strategy is to press the objector to justify his confidence that all functional analyses, however complicated, are systematically vulnerable to the absent qualia argument. In describing the absent qua-

lia argument, Block and Fodor present it as a general objection to functional accounts. However, I think their contention that, "for all we know, it may be nomologically possible for two states to be functionally identical (that is, to be identically connected with inputs, outputs and successor states), even if only one of the states has qualitative content," should be understood in light of the epistemic qualification it contains.¹⁵ Given what we now know, including our knowledge of proposed functional accounts, we cannot dismiss the possibility of absent qualia cases. This does not show that we would be unable to do so if provided with more sophisticated functional accounts, any more than the weaknesses of Wegener's theory of continental drift showed that no sophisticated theory of continental drift could be right, or the shortcomings of eighteenth century kinetic theories of heat showed that heat could not be explained by a more sophisticated kinetic theory.¹⁶ What these two cases, and many others, indicate is that straightforward extrapolation from the failings of early versions of a theory to the inevitable failure of later versions is unwarranted.

We are now prepared to tackle the issue of functional analyses. I suggest that we may view functional analyses as more or less careful attempts to codify presently available traditional psychological classifications. My claim is not that philosophers have regarded their functional accounts in quite this way. Their stated aim has usually been to produce accounts of the states we refer to using mental terms solely by reference to functional relationships. However, declarations of this intention have often been accompanied by claims that ordinary psychological classifications work solely or primarily through functional roles. Author's intentions aside, regarding them as codifications of traditional psychological classifications gives functional accounts two important roles to play in the psychological theory. Without functional analyses, it would be virtually impossible to state the theory. There would be no point of reference for the claim that, in using mental terms, we refer to states which can be characterized "psychologically." If functional analyses had not already shown what traditional psychological characterizations look like when made explicit, this step would need to be taken before it would be possible to explain what is meant by a "psychological" classification.

More importantly, if the psychological theory is right, functional analyses are what will be extended to produce an adequate classification of mental states. These analyses will turn out to have been preliminary sket-

ches for an adequate classification of mental phenomena. I think philosophers of mind have regarded functional analyses as important, partly because they feel that mental states are best understood primarily in terms of functional inter-relations. For a psychological theorist, functional analyses would still be important for exactly this reason.

In conclusion, I suggest that the positions I have presented as the “physiological theory” and the “psychological theory” probably should not be taken to be genuinely new versions of the identity theory. Several factors point to the conclusion that if central state materialists abandon the descriptive theory of reference in favor of a causal theory, the result will be the physiological theory; and that the psychological theory is just functionalism supported by a different theory of reference. As I read them, the main thrust of Smart’s and Armstrong’s efforts on behalf of materialism has been to make it reasonable to say that we refer to physical states in using mental terminology. Certainly Smart’s “topic neutral translations” were intended only to show how mental terms could refer to physical states.¹⁷ Although Armstrong offers numerous, interesting causal analyses of mental “concepts,” the analyses are not ends-in-themselves. Their purpose in the program is to allow the hypothesis that mental states are brain states to be a meaningful suggestion, while avoiding the pitfalls encountered by Place and Smart.¹⁸

By contrast, functionalists have not been concerned to show how mental terms could refer to states which are actually physical, but rather with defending the claim that mental states are properly characterized solely or primarily in terms of functional roles. Putnam and Fodor both claim that the assertion that mental states are physical states is really quite misleading. The difficulty with the claim is that it suggests that whenever we are in mental states of a certain type, we are in a particular type of physical state. They contend that, while each instance of a mental state is, of course, some physical state, it is wildly implausible to suppose that physiological classifications will turn out to be congruent with mental classifications.¹⁹ According to Putnam and Fodor, the states we refer to in using a particular mental term will not all be classed together as a certain kind of physical state, but they can all be captured by the same psychological or functional characterization. This is very close to the position maintained by psychological theorists. Again, I will conjecture that functionalists who do not have independent commitments to particular views of language will find

the psychological theory to be a more defensible version of their basic position.²⁰

A final reason for assimilating the physiological theory to central state materialism and the psychological theory to functionalism is the way the theories line up on difficult cases, such as those of Ernie and Bert discussed in Section I. Ernie is in a state classified by physiology as “ABC.” We assumed that ABC-states typically follow bodily injury, precede crying or groaning and so forth, but that, in this instance, the usual causal connections are absent. Bert is in physiological state “DEF,” a kind of state which almost never occupies the functional role of pain. Through some odd circumstances, however, Bert’s present DEF state has the typical causes and effects of pain. Physiological theorists believe that the states we refer to as “pains” will turn out to be particular kinds of physiological states. The claim that we have been talking about these states all along will be justified by the existence of correlations between our “pain” utterances and the states. In the hypothetical cases, our “pain” utterances would be systematically correlated with ABC-states, but not with DEF-states. Thus, like the central state materialist, a physiological theorist would take Ernie, but not Bert, to be in pain. If the psychological theory is right, then, eventually, presently available functional analyses will be extended to produce identifying descriptions of mental states. The psychological classification of pains would be some elaboration of: “states which follow bodily injury, precede crying and groaning and induce a desire that they stop.” Thus, like functionalism, the psychological theory would take Bert, and not Ernie, to be in pain.

I have tried not to favor either the physiological or the psychological theory. In this paper, my goals have been to show what the two main versions of the identity theory look like when recast in terms of a new theory of reference, and to argue that both versions can be defended against the classic arguments about the qualitative character of sensations that have plagued the identity theory since its inception. If they adopt the strategy I propose, all central state materialists and functionalists have to say about sensations and qualia is that they lack an account of qualitative character because, to date, no complete theory of sensations has been produced.²¹

NOTES

¹ Cf. D. M. Armstrong, *A Materialist Theory of Mind* (New York: The Humanities Press), 1968, p. 83.

² My presentation of the difference between functionalism and central state materialism derives partly from David Lewis' discussion in 'Mad Pain and Martial Pain,' in N. J. Block (ed.), *Readings in the Philosophy of Psychology*, Vol. I (Harvard University Press) pp. 216-222.

³ 'What Psychological States Are Not,' *The Philosophical Review* 81 (1972), 159-181, p. 173.

⁴ *Philosophical Studies* 27 (1975), 291-316.

⁵ *Ibid.*, p. 297.

⁶ This question is importantly ambiguous. It could mean: how do we use mental terms to refer to physical kinds? or, how do we use mental terms to refer to states, each of which belongs to some physical kind. Strictly speaking, an identity theorist need answer only the second, weaker question (see my definition, p. 1). See below, pp. 14-16.

⁷ Hilary Putnam has attacked the thesis that terms refer to objects by being associated with an identifying description of the objects in several important papers including, 'Is Semantics Possible?' *Metaphilosophy* 3 (1970), 187-201; 'Meaning and Reference,' *The Journal of Philosophy* 70 (1973), 699-711; and 'The Meaning of "Meaning",' in *Language, Mind, and Knowledge (Minnesota Studies in the Philosophy of Science)*, vol. VII, edited by Keith Gunderson (University of Minnesota Press), 1975, pp. 131-193. Saul Kripke defends a very similar view in 'Naming and Necessity,' in D. Davidson and G. Harman (ed.), *Semantics of Natural Language* (Dordrecht: D. Reidel), 1972, pp. 253-255. Keith Donnellan attacks the descriptive theory of reference as an account of proper names in several papers, including 'Proper Names and Identifying Descriptions,' also in Davidson and Harman (eds.), pp. 356-379 and 'Speaking of Nothing,' *The Philosophical Review* 82 (1974), 3-31.

⁸ Cf. 'Speaking of Nothing,' *op. cit.*

⁹ I attribute to Putnam the first "ability account" of consistent linguistic usage on the basis of such remarks as "How do I convey the meaning of the word 'lemon'? Very likely I show the man a lemon." (from 'Is Semantics Possible?' *op. cit.*, p. 58). Michael Devitt discusses this view explicitly in connection with proper names in 'Singular Terms,' *The Journal of Philosophy* 71 (1974), 183-205. Putnam discusses "stereotypes" in 'Is Semantics Possible?' and 'The Meaning of "Meaning",' *op. cit.* cf. pp. 58-63 and 147, 169-173, respectively.

In these passages Putnam's aim is to reveal what is behind the traditional philosophical notion of "meaning." However, I think his remarks about conveying the meanings of words implicitly contain the solutions to questions of consistent linguistic usage presented in the text.

¹⁰ I explore some of the reflections that lead philosophers to think that qualia are peculiarly inexplicable in 'Phenomenal Qualities,' *American Philosophical Quarterly*, April 1979, 123-129.

¹¹ In 'Naming and Necessity', *op. cit.*

¹² For the purposes of argument, in addition to his technical machinery, I accept Kripke's view that each sensation has a particular phenomenal quality, that is, that it makes sense to talk about the exact qualitative character of my last headache. In fact, I do not think this way of talking is justified, for reasons I present in 'Phenomenal Qualities,' *op. cit.*

Alternative accounts of the apparent contingency of the claims of the identity theory are offered by Michael E. Levin in 'Kripke's Argument against the Identity Thesis,' *The Journal of Philosophy* 72 (1975), 149-167, and by George Sher in 'Kripke, Cartesian Intuitions, and Materialism,' *The Canadian Journal of Philosophy* 7 (1977), 227-238.

¹³ For the sake of completeness, I will note that it is possible to take two further positions about the usefulness of physiological and psychological classifications of mental states. Besides the psychological theory and the physiological theory, an identity theorist could endorse psychological classifications and be neutral about the utility of a physiological taxonomy, or he could support the latter and be neutral about the former.

¹⁴ Some philosophers may resist granting the label "law" to such claims as "pains typically induce a desire that they cease." The objection could be that current psychological generalizations are too vague or, perhaps, too platitudeous to be laws. This is an enormous issue which I cannot tackle. All I wish to claim is that, since common psychological principles are universal in form, capable of supporting counterfactuals, projectable from their instances, and together provide a standard method of predicting and explaining a range of phenomena, it is reasonable to view them as laws until shown otherwise.

¹⁵ *Loc. cit.*

¹⁶ Cf. A. Hallam, *A Revolution in the Earth Sciences*, Oxford, The Clarendon Press, 1973, and J. B. Conant, editor, *Harvard Case Histories in Experimental Science*, Harvard University Press, 1948, Vol. 1, case 3, 'The Early Development of the Concepts of Temperature and Heat: The Rise and Decline of the Caloric Theory,' prepared by Duane Roller, especially pp. 150-155.

¹⁷ Cf. J. J. C. Smart, 'Sensations and Brain Processes,' reprinted in *Materialism and the Mind-Body Problem*, edited by David Rosenthal (Prentice-Hall), 1971, pp. 53-66.

¹⁸ Cf. Armstrong, *op. cit.*, pp. 82 and 89.

¹⁹ Cf. Putnam's 'The Nature of Mental States' and Fodor's 'Materialism,' both reprinted in Rosenthal, *op. cit.*, pp. 150-161 and 128-149, respectively.

²⁰ One philosopher, who would be classified as a functionalist on my way of dividing the field, and who has independent commitments in philosophy of language, is David Lewis, Cf. 'How to Define Theoretical Terms,' *The Journal of Philosophy* 67 (1970), 427-438 and 'Psychophysical and Theoretical Identifications,' *Australasian Journal of Philosophy* 50 (1972), 249-258.

²¹ A number of people have given me useful comments on earlier drafts. I am particularly indebted to Philip Kitcher. Part of the work was completed while I enjoyed a Summer Institutional Grant from the University of Vermont.

Manuscript submitted 24 March 1981
Final version received 10 August 1981

A BRIDGE BETWEEN COGNITIVE SCIENCE AND NEUROSCIENCE: THE FUNCTIONAL ARCHITECTURE OF MIND

(Received 20 September, 1982)

Discussion within philosophy on the relation of cognition to neural phenomena, or of mind to body, has focused on two views in recent decades. Both views accept that mental events themselves are brain events with physical descriptions. But whereas the Identity Theory holds out the hope of nomological connections between types of brain events and types of mental events, Functionalism denies any such relation. Identity theorists envision an ultimate reduction of cognitive psychology to neuroscience whereas most functionalists deny that possibility. (Australian Identity Theorists such as Smart, 1959, and Armstrong, 1968, present that Identity Theory as a version of Functionalism. Smart's topic-neutral specification of psychological events, for example, specifies mental events in terms of their interactions with other mental events or with sensory stimuli or behaviors. When I speak of Functionalism, however, I shall be referring to the American variety which sees Functionalism as an alternative to the Identity Theory.) This paper proposes that we approach the relation of mental events to brain events in a quite different way, using Pylyshyn's (1980) notion of a functional architecture. After an overview of the conflict between Identity Theory and Functionalism, I will turn to explicating the notion of a functional architecture and show how it provides a more useful ontological framework for understanding how neuroscience and the emerging discipline of cognitive science can relate to each other.

1. THE ALTERNATIVES OF IDENTITY THEORY AND FUNCTIONALISM

In his classical defense of the Identity Theory, Smart (1959) begins by accepting the scientific perspective that there is "nothing in this world but increasingly complex arrangements of physical constituents" and seeks to avoid having to admit sensations as peculiar "nomological danglers" to this perspective by viewing them as "brain processes of a certain sort," so that

“sensations are nothing over and above brain processes.” To do this, Smart introduces the notion of strict identity — the events that were taken to be two are in fact the same. Smart does not defend his position by a direct argument but by dialectically showing that the traditional objections to Identity Theory are conceptually confused. For example, he argues that even though the events may be thought of differently when one describes them in mental terms from when one describes them in physical terms, the events could still be the same, just as Venus is the same whether thought of as the morning star or the evening star. Since the objections fail to undermine the Identity Theory, Smart claims that “principles of parsimony and simplicity” tell for it against dualistic epiphenomenalism.

Although Smart does not explicitly make the claim, his motivation for defending Identity Theory (the scientific peculiarity of nomological danglers) would suggest that he would see Identity Theory as providing a way in which neuroscience could explain mental events. By virtue of their identity with neural events and the laws accounting for the interactions of neural events, one would explain the occurrence of mental events. In viewing neural events as explaining mental events, two possibilities present themselves. The first is a type-type identity of neural events and mental events where a particular type of neural event would always also constitute a particular type of mental event. Such a situation would allow for universal bi-conditional statements relating mental and physical descriptions of the event and so would certainly permit a Nagel (1961) style reduction of psychological laws that refer to events under their mental descriptions to neuroscience laws that refer to the same events under their physical descriptions. This seems to be the treatment of identity most congenial to Smart’s objectives in defending the Identity Theory, since it clearly eliminates the need for either dangling events or laws.

An alternative to the type-type treatment of identity is a token-token view of identity according to which each token of a mental event type is identical with a token of a physical event type, but where each mental event type is not projectable onto a single physical event type. This is the perspective of many Functionalists who defend the non-projectability of mental event types onto physical event types by arguing for a many-many mapping of mental descriptions of events onto physical descriptions. Putnam (1975) argues that we attribute the same beliefs and desires to different people despite the fact that the best neurophysiological theories point to differences between persons’ brains. Moreover, he claims that totally alien beings with very different

brain structures could also possess these mental states. Fodor (1975) presses the reverse claim that the same type of neural state could be the instantiation of different mental states. In both cases, a computer analogy is used to make the claims seem plausible. The same program can be run on different actual machines while the same physical state of the machine may occur in the course of running very different programs.

The rejection of type-type identity theory is taken by Putnam and Fodor to show that psychological explanations are not reducible to neuroscience ones. Without biconditionals linking mental and physical descriptions, no Nagel type reduction is thought to be possible. (To this claim Fodor and also Pylyshyn, 1980, add the claim that it is the psychological and not the neurological laws that are needed to explain things like human behavior.)

Richardson (1979 and 1982) has countered this argument. Nagel type reductions do not require biconditionals — only the ability to specify the circumstances under which an event satisfying a particular physical predicate will satisfy the corresponding mental predicate. As long as these conditions can be specified in the correspondence rules, one could produce the proper derivation of psychological laws from physical laws.

Even if Functionalists accepted Richardson's claims, however, there is still a crucial distinction they could draw between their position and that of an Identity Theorist who accepted type-type identities. The conditions under which an event satisfying a neural description also satisfied a mental description would typically be specified in terms of other events described in psychological terms. The point can best be understood by using the computer analogy — to tell whether a particular computer state satisfied a program statement would require specifying the program of which it is a part. Only in the context of running a certain program does a particular electrical activity in the hardware count as adding. Likewise only in the context of specific other mental states does a brain state constitute a particular mental state. Although one might also be able to go through these other mental states and specify them in terms of their neural state so as to remove mental terms from the correspondence laws, one has still embedded into the correspondence laws all the regularities typically stated in psychological terms. One has not, then, eliminated psychological laws by reducing them to physical laws, but has set out the conditions under which events satisfying a physical description also satisfy a mental description. The correspondence laws are doing the work of the 'reduction', not the physical laws. Thus, the difference

between the type-type Identity Theorist and the Functionalist is that, for the former, physical laws end up explaining mental phenomena, while for the Functionalist one still needs to specify separately the conditions under which the events referred to in physical laws satisfy mental descriptions.

2. THE NEED FOR AN ALTERNATIVE MODEL

For the most part, I accept the arguments of Putnam and Fodor against the Identify Theory, at least as I have recast them above. But I find the framework in which the controversy between the Identity Theory and Functionalism is set to be a poor one for understanding the relation between neuroscience and cognitive science. One reason is the emphasis put on the notion of formally stated laws. Thinking about the laws of a science rather than the phenomena described in a science has become traditional within analytic philosophy of science, but that may be inappropriate in areas like neuroscience and psychology where, for the most part, the objectives of research are not formally stated laws but models of how a system works. Within the context of a model one is not so inclined to worry about how to translate between different descriptions of the system's operation but to consider how different descriptions provide different information about the operation of the system.

When one appeals to a Nagel type model of reduction, one usually assumes that the terms of the different theories refer to the same entities and that bridge laws relate terms referring to the same entity. But when one looks at a complex system, one realizes that one can enumerate the parts in different ways. The parts enumerated in one way may themselves be assemblies of parts enumerated in other ways. One may have theoretical accounts or models of the interaction of either sets of these parts. The correct way to connect these theoretical accounts would not be by identifying the terms referring to parts in both theories (since they do not refer to the same parts) but by explaining how the parts as described by one theory form the assemblies described in the other theory. This involves a composition relation, not an identity relation. (Cf. Wimsatt, 1976, for some suggestions as to the ontological framework in which different levels are related by part-whole relations.)

Although a composition relation between structural parts is perhaps the easiest to conceptualize, there is another type of composition relation that is perhaps more relevant for discussing the relation of mental events to neural

events. This is the composition relation between processes in a system. Cummins (1975) presented the idea of a functional analysis as one kind of explanation of how a process is performed. His proposal is that one analyze the operation of a system into component processes and their interactions and that one repeat this process until one reaches such basic processes that one sees how to provide a traditional covering law explanation of how they are performed by structures within the system. (Dennett's 1978 account of the relation of a design stance and physical stance explanation of a system is essentially similar.) Within Cummin's type of functional analysis, one can have several embedded functional accounts — a process may contribute to and thus help explain another process and yet be explained in terms of its component processes.¹

Composition relations between processes are common in biology. The process of an enzyme interacting with a substrate may be part of the process of glycolysis, which is itself part of the process of energy metabolism that is, in turn, part of the process of muscle action. Different levels of inquiry focus on each of these processes — enzyme biochemistry, pathway biochemistry, bioenergetics, and muscle physiology. There is a close interaction between these inquiries — what is known about enzyme activities ought to explain how they can figure in the processes described by pathway biochemistry. In connecting these inquiries, however, one's attention is not on finding identities between processes but on figuring out how the performance of one process depends on or facilitates another. Insofar as one would speak of a reductive explanation, it would refer to this attempt to explain how a process results from putting together sub-processes in a particular manner. In the next section I will sketch a framework for relating neuroscience to cognitive science that will involve a composition relation between processes.

3. THE NOTION OF A FUNCTIONAL ARCHITECTURE

In learning to program a computer, one learns a programming language like LISP, FORTRAN, or BASIC. Such a language is what specifies the basic operations that one can get the computer to perform. Most often the language one learns in computer programming is not the basic operating language of the machine. Through either an interpreter or a compiler one's programming language is translated into a set of machine language commands that the machine then performs. The machine language is able to direct the behavior

of the computer because its symbols are directly mapped onto operations in the hardware of the machine. This mapping of symbols in the programming language onto hardware operations defines what Pylyshyn (1980) calls the "functional architecture" of the machine.

Because of the use of interpreters and compilers with computers, the definition of the functional architecture is not absolute. For the programmer, it is the indirect mapping of the higher level programming language that provides the programmer with the capacities he or she can employ. The programming language then constitutes what Pylyshyn refers to as a "virtual machine." When one is using one computer to simulate the operation (not just the output performance) of another, it is assumed that the machines may not have the same ground level functional architecture, but one must at least be able to specify a level at which they constitute the same virtual machine. Only when the two machines share a common set of capacities does it make sense to say that they utilize these capacities in the same way. When, in what used to be referred to as the "cognitive simulation" approach to artificial intelligence research, one tries to simulate human information processing, one also needs to be able to specify a level at which the human and the computer constitute the same virtual machine. In the case of humans, Pylyshyn assumes that there is no interpreter or compiler function employed – the programming language is the operating language of the brain. (This assumption is ultimately not crucial. If it is false, however, the picture becomes much more complex.) Then, in order to simulate human information processing on a computer, one must discover the functional architecture of the human being. In the case of humans, the functional architecture will consist in a mapping of neural states onto symbolic expressions.

Since, in order to use a computer to study the steps in human information processing, one needs to identify the processing capacities humans use and build those into the computer (possibly through an interpreter or compiler), Pylyshyn (1981) makes the task of discovering the functional architecture of the human mind a high priority for cognitive science. But, while construing the architecture as defined by a mapping from neural states onto symbolic expressions, Pylyshyn largely discounts neuroscience from having a role to play in identifying the functional architecture.² Rather, he proposes a functional criterion for discovering the functional architecture of the human mind. This functional criterion is made possible by the fact that it is only in terms of a mapping of neural states onto expressions that the brain can process the

information carried in a particular set of expressions. Pylyshyn argues that the mapping cannot itself be altered by supplying information to the system in symbolic form. Thus, the functional architecture is impenetrable to information. So, by identifying information processing operations that cannot be altered by information, Pylyshyn proposes that we can identify the functional architecture itself.

After developing the cognitive impenetrability criterion for belonging to the functional architecture, Pylyshyn (1980) is unable to find many processes that are cognitively impenetrable. For him, then, the task of discovering the functional architecture, a necessary condition for using computers to study human information processing, has barely begun. Although the concept of a functional architecture has yet to pay fruitful dividends as Pylyshyn applies it, I contend that it has a significant role to play. Since he introduced the notion of the functional architecture as a bridge between the neural or hardware level and the mental or programming level, it can provide a framework for integrating neuroscience and cognitive science research. Ultimately, if neuroscience and cognitive science research is integrated at the point of the functional architecture, one may acquire the requisite knowledge of the functional architecture that is required for computer models to be truly informative about the processes of human information processing.³

The virtue of the notion of a functional architecture is that it does not seek to identify processes at two levels of inquiry but provides a compositional relation between processes at these levels. Through the functional architecture, a set of neural processes is mapped onto a symbol which then figures in the processes studied in information processing models. Within this framework, cognitive science theories are not just translations of neuroscience theories – rather, they describe the interactions of components whose operations are, in turn, studied by neuroscience.

At this point, the following objection is likely to be forthcoming: why should neuroscience be limited to studying the component processes that get mapped onto symbols and not be allowed to study the interaction of those component processes themselves? This is a fundamental question and, by trying to answer it in the remainder of this paper, I will try to show more of what is involved in the proposal that the functional architecture is the bridge between neuroscience and cognitive science.

The objection proposes that neuroscience can go beyond component processes in the brain to itself study the interaction of these processes. What I

want to argue is that even though this further study may involve studying brain mechanisms and may utilize some of the same techniques as are used in studying the component processes, this investigation may not be merely an extension of the study of component processes. The reason is that this study will involve looking at how these components are organized so as to interact with each other in selected ways. The organization of components gives rise to a higher level of organization. The best way to understand this point about levels of organization in the brain is through an analogy to how levels of organization affect physiological inquiry. Physiological phenomena are ultimately the product of chemical activities. Yet, physiology adds something to chemical studies — a focus on how chemical entities are organized in the system. It was Claude Bernard (1865) who recognized the important role of organization in physiology. He argued that conditions within the organism — in what he called the “internal environment” — could affect the responsiveness of particular components within the organism.⁴ The reason is that chemical reactions are environmentally sensitive and so can be regulated by controlling the environment in which they occur. The control over the internal environment is realized through the components in the system being organized to achieve such effects as negative feedback. The higher level thus regulates the lower level so that we realize what Campbell (1974) called “top-down” causation.⁵

To apply this analogy to the case of the mind, I am proposing that the component neural processes are designed (by evolution) to be capable of information processing. Once the system is engaged in information processing, the regularities of information processing (e.g., the principles of thought) govern the behavior of the individual components (i.e., the symbols that are mapped onto neural processes) in much the way principles of negative feedback govern the chemical reactions occurring in a living organism. These principles are what cognitive psychology is seeking to discover. Insofar as they play a role like negative feedback processes play in physiology, they affect the behavior of the component neural operations. Cognitive psychology thus seeks to discover a set of regularities that are not the same as or the sum of the neural processes operating in individual brain mechanisms. These regularities therefore cannot be eliminated in a complete science of the mind.

It should be possible (at least in specific cases) to describe these regularities at the level of their operation on individual brain components. If this is what is meant when the objector proposes that neuroscience should be allowed to

extend beyond the study of component mechanisms, that view is not inconsistent with the one offered here. My contention is, though, that it is really the principles of cognitive science — the information processing procedures by which symbols are manipulated — that neural scientists are studying when they are considering how the component processes are being affected by being incorporated in a whole brain (and ultimately in a person who interacts with an environment). The computer scientist could similarly study the transformations occurring in the hardware of a computer when it is processing information. Further, the chemist can study the reactions occurring at enzyme sites as a result of negative feedback that results from the body's organization. In each case, however, the principles that are operative are those governing the interactions of the components and these are studied by the higher level science.

Drawing a parallel to yet another area of multi-level scientific inquiry may be useful at this point. In the decade after the rediscovery of Mendel's work, Mendelian genetics was viewed as an alternative to Darwinian evolutionary theory. With the advent of population genetics, the two disciplines were linked. Mendelian genes became viewed as the hereditary mechanisms underlying evolution. However, at the same time, a reductionistic research bias was set in which it was attempted to understand evolution as a process occurring at the gene level (cf. Williams, 1966, and Dawkins, 1976). While clearly the gene level can be used as a 'book-keeping' level at which one records the changes in gene frequencies due to selection, it may not be the appropriate explanatory level. To understand why gene frequencies change, one must look at the units that do interact with an environment (individual organisms, kin or social groups, populations) and study which of their properties are interacting with environmental factors to change gene frequencies. A reductionistic bias (Wimsatt, 1980) has led population genetics researchers to ignore this level and to settle for summarizing its effects. But it is to levels at which genes are only components of larger wholes that one must turn to understand changes in gene frequency. (For examples of how higher levels such as that of groups may need to come into play in genetics, see Wade, 1978, and Wilson, 1980.)

My proposal, then, is to think of information processing as a higher level process involving the interaction of the individual neural processes that fall in the special domain of neuroscience. Those neuroscientists who focus on the interactions of neural components are studying information processing. This

fits with the fact that neuroscientists who do investigate higher level connections in the brain do try to describe these processes in cognitive terms of learning, memory, etc. (See, for example, Kandel, 1978; Thatcher and John, 1977; and O'Keefe and Nadel, 1978.) These investigations are looking at a level beyond the individual neuron or ganglia to their interactions and are tracing the flow of information through the brain. They are working on the same side of the functional architecture as cognitive scientists who try to model how the brain manipulates symbols.

Thus, in the case of neuroscience and cognitive science, I am proposing that it is the functional architecture that provides the bridge between the components themselves and the higher level processes. It is at the point of the functional architecture that component processes come to form the types of entities whose behavior cognitive science explores. This compositional account of mind/brain relations seems to give a more useful handle on the relation of cognitive science to neuroscience than the identity relation employed in either the Identity Theory or Functionalism in that it shows how the two inquiries can complement each other.⁶ Within this framework one is no longer proposing to reduce and replace cognitive science with neuroscience, but only seeking to show how neural processes are the constituents in cognitive operations.

NOTES

¹ Cummins' account of functional analysis is important in freeing one from allegiance to the covering law model of explanation and leading one to look at causal processes. As an analysis of functional statements, however, it risks being too general since any process resulting from causal interaction can be given a causal explanation on Cummins' account. Insofar as it focuses just on the pattern of causal interaction, a functional analysis of this type loses content. Thus, Block (1978) charges that Functionalism provides too liberal a characterization of mental states unless it incorporates a reference to the physical matter of the human brain, in which case it becomes too 'chauvinistic'. But there is another alternative. In biology, not every causal effect of an organ is counted as a function; only those which enable the organ to fulfill some greater role. Wimsatt (1972) presents an analysis of function statements that makes reference to the purpose of a process. This purpose may be stipulated in a non-arbitrary way by appealing to the selection process through which the entity came to exist. Wimsatt's analysis of function thus leads one to focus on the context in which an operation is being performed. Richardson (forthcoming; see also Lycan, 1981, and Sober, forthcoming) argue that through using Wimsatt's analysis of function one can properly constrain the functional analysis of mental events to ones truly mental and so avoid Block's objection.

² Pylyshyn acknowledges a possible role for neuroscience in discovering the functional architecture of the mind, but he is not strongly supportive of pursuing such endeavors: "Although the information-processing approach deliberately sets a level of analysis

independent of specific material forms, it cannot entirely ignore physiological considerations – particularly as a source of evidence for potentially constraining the structure of the processes (although this source of constraint has been remarkably impotent in the past)" (1981, p. 90).

³ In Bechtel (forthcoming) I discuss a number of examples of research in neuroscience that appear to be informative about the functional architecture of the mind including the recordings of macropotentials from the scalp, the localization of functions in cortical structures, and the investigation of non-linear, non-digitalized transmission mechanisms, and I argue that neuroscience does have the resources for showing that the brain has a non-von Neumann architecture.

⁴ Bernard proposed to use this regulatory capacity of the internal environment to answer the kinds of arguments Bichat had made for vitalism. I explicate Bernard's arguments in Bechtel (1982a) and propose that by following Bernard's strategy materialists will be able to answer a number of arguments commonly made on behalf of dualism.

⁵ One of the important features of the effect of higher level systems on lower level components is to change the behavior of the lower level components. In Bechtel (1982c), I argue that this makes it impossible to ignore the effects of the higher level systems in studying lower level processes or to complete a lower level science without knowing all the higher level regularities. Given the possibility of a virtually unlimited number of upper level organizational patterns, I argue that the scientific enterprise may be incompletable.

⁶ I have previously argued (Bechtel, 1982b) that one of the advantages of linking the inquiries of different disciplines together is that even if the links are themselves fallible, proposing the connections increases the possibility of detecting errors in one of the fields of investigation. The theories in each discipline become constrained by what is discovered about the phenomena investigated in the other discipline and the additional constraints help reveal errors that may exist in the explanations originally offered.

BIBLIOGRAPHY

- Armstrong, D. M.: 1968, *A Materialist Theory of Mind* (Humanities Press, New York).
- Bechtel, W.: 1982a, 'Taking vitalism and dualism seriously: toward a more adequate materialism', *Nature and System* 4, pp. 23–43.
- Bechtel, W.: 1982b, 'Two common errors in explaining biological and psychological phenomena', *Philosophy of Science* 49, pp. 549–574.
- Bechtel, W.: 1982c, 'Forms of organization and the incompleteness of science', presented at conference on *Limits of Scientific Knowledge*, University of Pittsburgh Center for the Philosophy of Science (publication forthcoming).
- Bernard, C.: 1865, *Introduction à l'étude de la médecine expérimentale* (Baillière, Paris). English translation by H. C. Greene: 1959, *An Introduction to the Study of Experimental Medicine* (Dover, New York).
- Block, N.: 1978, 'Troubles with functionalism', in *Perception and Cognition: Issues in the Foundations of Psychology* (Minnesota Studies in the Philosophy of Science, Volume 9), ed. Wade Savage (University of Minnesota Press, Minneapolis), pp. 261–325.
- Campbell, Donald T.: 1974, 'Downward causation in hierarchically organized biological systems', in: *Studies in the Philosophy of Biology*, ed. F. J. Ayala and T. Dobzhansky (University of California Press, Berkeley).
- Cummins, R.: 1975, 'Functional analysis', *The Journal of Philosophy* 44, pp. 43–64.
- Dawkins, R.: 1976, *The Selfish Gene* (Oxford University Press, Oxford).

- Dennett, D. C.: 1978, *Brainstorms* (Bradford Books, Montgomery, Vermont).
- Fodor, J. A.: 1975, *The Language of Thought* (Crowell, New York).
- Kandel, R.: 1979, 'Small systems of neurons', *Scientific American* 238, pp. 66-76.
- Lycan, G.: 1981, 'Form, function, and feel', *The Journal of Philosophy* 78, pp. 24-49.
- Nagel, E.: *The Structure of Science* (Harcourt and Brace, New York).
- O'Keefe, J. and Nadel, L.: 1978, *The Hippocampus as a Cognitive Map* (Clarendon Press, Oxford).
- Putnam, H.: 1975, *Mind, Language, and Reality: Philosophical Papers*, Volume 2 (Cambridge University Press, Cambridge).
- Plyshyn, Z.: 1980, 'Computation and cognition: issues in the foundation of cognitive science', *Behavioral and Brain Sciences* 3, pp. 111-132.
- Plyshyn, Z.: 1981, 'Complexity and the study of artificial and human intelligence', in: *Mind Design*, ed. J. Haugeland (Bradford Books, Montgomery, Vermont). This is a revised version of a paper that first appeared in *Philosophical Perspectives on Artificial Intelligence*, ed. M. Ringle (Humanities Press, Atlantic Highlands, N.J., 1979).
- Richardson, R. C.: 1979, 'Functionalism and reductionism', *Philosophy of Science* 46, pp. 533-558.
- Richardson, R. C.: 1982, 'How not to reduce a functional psychology', *Philosophy of Science* 49, pp. 125-137.
- Richardson, R. C.: forthcoming, 'Top down strategies, reductionist heuristics and localization of function'.
- Smart, J. J. C.: 1959, 'Sensations and brain processes', *Philosophical Review* 68, pp. 141-156.
- Sober, E.: forthcoming, 'Putting the function back into functionalism'.
- Thatcher, R. W. and John, E. R.: 1977, *Foundations of Cognitive Processes* (Lawrence Erlbaum Associates, Hillsdale, N.J.).
- Wade, M. J.: 1978, 'A critical review of the models of group selection', *Quarterly Review of Biology* 53, pp. 101-114.
- Williams, G. C.: 1966, *Adaptation and Natural Selection* (Princeton University Press, Princeton).
- Wilson, D. S.: 1980, *The Natural Selection of Populations and Communities* (Benjamin, Menlo Park, Cal.).
- Wimsatt, W. C.: 1972, 'Teleology and the logical structure of function statements', *Studies in the History and Philosophy of Science* 3, pp. 1-80.
- Wimsatt, W. C.: 1976, 'Reductionism, levels of organization, and the mind-body problem', in G. Globus, G. Maxwell, and I. Savodnik (eds.): *Brain and Consciousness* (Plenum Press, New York).
- Wimsatt, W. C.: 1980, 'Reductionistic research strategies and their biases in the units of selection controversy', in Tom Nickles (ed.): *Scientific Discovery*, Volume 2: *Case Studies* (D. Reidel, Dordrecht), pp. 213-259.

*Department of Philosophy,
Georgia State University,
University Plaza,
Atlanta, GA 30303,
U.S.A.*

EPILOGUE

CONFLICTING CONCEPTIONS

LANGUAGE AND MENTALITY:
COMPUTATIONAL, REPRESENTATIONAL, AND
DISPOSITIONAL CONCEPTIONS*

(A) cognitive theory seeks to connect the *intensional* properties of mental states with their *causal* properties *vis-à-vis* behavior. Which is, of course, exactly what a theory of the mind ought to do.

Jerry Fodor

ABSTRACT. The purpose of this paper is to explore three alternative frameworks for understanding the nature of language and mentality, which accent syntactical, semantical, and pragmatical aspects of the phenomena with which they are concerned, respectively. Although the computational conception currently exerts considerable appeal, its defensibility appears to hinge upon an extremely implausible theory of the relation of form to content. Similarly, while the representational approach has much to recommend it, its range is essentially restricted to those units of language that can be understood in terms of undefined units. Thus, the only alternative among these three that can account for the meaning of primitive units of language is one emphasizing the basic role of skills, habits, and tendencies in relating signs and dispositions.

There are several reasons why the nature of language and mentality is fundamental to research in artificial intelligence and to cognitive inquiry in general. One tends to be the assumption – better viewed as a *presumption* – that thinking takes place in language, which makes the nature of language fundamental to the nature of mental processes, if not to the nature of mind itself. Another is that computers operate by means of software composed by means of a language – not a natural language, to be sure, but a computer language, which is a special kind of artificial language suitable for conveying instructions to machines. And another is that debates continue to rage over whether or not machines can have minds, a question whose answer directly depends upon the nature of mentality itself and indirectly upon the nature of language – especially the nature of languages suitable for use by machines.

Below the surface of these difficulties, however, lies another problematic question, namely: is artificial intelligence *descriptive* or *normative*? For if artificial intelligence is supposed to utilize the methods that human beings themselves – descriptively – actually employ in problem solving, then there

would appear to be a powerful motive for insuring that the languages used by machines are similar (in all relevant respects) to those used by humans. If artificial intelligence is *not* restricted to the methods that human beings actually employ but may utilize those that humans should use – normatively – whether or not they actually do, then whether computer languages are like or unlike natural languages at once appears to be a less pressing issue.

Most students of artificial intelligence tend to fall into two broad (but heterogeneous) camps. One camp maintains the ‘strong’ thesis that AI concerns how we do think. The other maintains the ‘weak’ thesis that AI concerns how we ought to think. And there are grounds to believe that the predominant view among research workers today is that the strong thesis is correct. Eugene Charniak and Drew McDermott, for example, envision AI as “the study of mental faculties through the use of computational models” [Charniak and McDermott (1985), p. 6]. An assumption that underlies this approach is that, at some level, the way in which the mind functions is the same as the way in which certain computational systems – digital computers, especially – also function. This assumption, however, is one that adherents of both camps might endorse, insofar as even normative approaches to AI presumably would need to satisfy this condition ‘at *some* (suitable) level’.

As though to disabuse those who might mistake the conception that they endorse for a normative one, Charniak and McDermott go so far as to assert that, “The ultimate goal of AI research (which we are very far from achieving) is to build a person, or, more humbly, an animal” [Charniak and McDermott (1985), p. 7]. Although their position may be extreme in this respect, much of the impetus for AI and cognitive science research along these lines arises from the *symbol system hypothesis* advanced by Alan Newell and Herbert Simon, according to which the necessary and sufficient conditions for general intelligence are those possessed by *physical symbol systems*, which are physical systems (or ‘causal systems’) that have the capacity to process/manipulate/... sequences of marks/signs/... from a designated vocabulary [cf. especially Newell and Simon (1976), pp. 40–42].

This general approach, moreover, has been reinforced by the proposition that, when mental processes are viewed as computational, minds themselves can be viewed as special kinds of formal systems. John Haugeland (1981), (1985), for example, has gone further in suggesting that mental activity can be adequately portrayed as the behavior of an *automated formal system*, a position that leads him to the conjecture, “Why not suppose that people *just are* computers (and send philosophy packing)?” [Haugeland (1981), p. 5]. Indeed, the prospect of reducing the philosophy of mind to problems of

design in the construction and development of digital machines has excited a host of adherents across many fields and disciplines, including those of linguistics, of psychology and of philosophy as well as those of computer science and AI.

The tenability of this computational conception, of course, has not gone unchallenged. In fact, from a perfectly general perspective rooted in (what is known as) *semiotic* (or 'the theory of signs'), there are three fundamental aspects to systems of signs, generally, and to languages, specifically. These distinctions, which were introduced by Charles S. Peirce and subsequently refined by Charles Morris (1938) and Rudolf Carnap (1939), concern, first, the relations that signs bear to other signs (known as 'syntax'); second, the relations that signs bear to that for which they stand (known as 'semantics'); and, third, the relations that signs bear to other signs, to that for which they stand, and to sign users (known as 'pragmatics'). After all, if there are three dimensions of signs, how could any theory of language that focuses on only *one* be expected to provide an adequate account of language and mentality?

The purpose of this paper, pursuing this lead, is to explore three alternative frameworks for understanding the nature of language and mentality, which accent syntactical, semantical, and pragmatical dimensions of the phenomena with which they are concerned, respectively. Although the computational conception currently exercises considerable appeal, its defensibility seems to hinge upon an extremely implausible theory of the relation of form to content. Similarly, while the representational approach has much to recommend it, its range is essentially restricted to those units of language that can be understood in terms of undefined units. Thus, the only alternative among these three that seems capable of accounting for the meaning of the primitive units of language emphasizes the role of skills, habits and tendencies relating signs and dispositions. And the same considerations provide a foundation for assessing conceptions of these kinds as 'models of the mind'.

Perhaps one cautionary note is in order before pursuing this objective. For the methodology to be employed here is analytical rather than historical, in the sense that the subject of investigation is the *problem space* – or, even better, the *solution space* – appropriate to these problems. I am less concerned with the detailed positions that have been held by the specific individuals – such as Haugeland, Fodor, Stich, and others – whose works are mentioned in passing than I am with the general features of the problems and solutions toward which they are directed. By emphasizing the predominantly syntactical, semantical, and pragmatical aspects of the alternatives considered, their essential dimensions may be perceived more clearly and their

relative plausibility may be assessed more accurately – a procedure not unlike relating surface phenomena to their deep structure.

1. THE COMPUTATIONAL CONCEPTION: SYNTACTICAL MODELS OF THE MIND

Computational conceptions of language and of mind depend upon the assumption that languages and mental processes can be completely characterized by means of purely formal distinctions. In discussing the computational conception (or ‘model’) of language and of mind, for example, Fodor has observed that such an approach entails the thesis that “...mental processes have access only to formal (non-semantic) properties of the mental representations over which they are defined” [Fodor (1980), p. 307]. Thus,

...the computational theory of the mind requires that two thoughts can be distinct in content only if they can be identified with relations to formally distinct representations. More generally: fix the subject and the relation, and then mental states can be (type) distinct only if the representations which constitute their objects are formally distinct. [Fodor (1980), p. 310].

Notice that at least two issues are intimately intertwined in this passage, for Fodor is maintaining (a) that thoughts are distinct *only if* they can be “identified” with distinct representations, without explaining how it is (b) that specific thoughts can be identified with specific representations. As a necessary condition for thought identity, in other words, condition (b) must be capable of satisfaction as well as condition (a); otherwise, his account will be *purely* syntactical as an analysis of the relations between signs lacking significance.

Fodor, Stich and others too have examined possible ways in which specific ‘thoughts’ might be identified with specific ‘representations’ (in other words, how forms could be infused with content). The strongest versions of the computational conception, however, tend to eschew concern for matters of semantics, as Stich, for example, as emphasized:

On the matter of content or semantic properties, the STM [the Syntactic Theory of the Mind] is officially agnostic. It does not insist that syntactic state types have no content, nor does it insist that tokens of syntactic state types have no content. It is simply silent on the whole matter ... the STM is in effect claiming that psychological theories have *no need* to postulate content or other semantic properties, like truth conditions. [Stich (1983), p. 186].

In order to preserve the differences between the syntactical character of the

computational conception and the semantical character of its representational counterpart, we shall defer our consideration of thesis (b) until Section 2.

Fodor is suggesting an exceptionally strong connection between the *form* of a thought and its *content*. The strongest version of such a position would appear to be that mental tokens (which might be sentences in a natural language, inscriptions in a mental language, or some other variety of types and tokens capable of formal discrimination) have the same content (or express the same thought, convey the same idea or otherwise impart the same information) *if and only if* they have the same form (where 'mental tokens', of course, are instances of mental types that might have content). Differences and similarities, formal or not, no doubt presuppose a point of view, which establishes a standard of 'sameness' for tokens with respect to whether or not they qualify as 'tokens of the same type'. Assuming that this condition can be satisfied by systems for which formal distinctions are fundamental, let us draw some examples from ordinary English as helpful illustrations.¹

Any biconditional, of course, is logically equivalent to the conjunction of two conditionals. In this case, those two conditionals are (i) if mental tokens have the same form, then they have the same content; and (ii) if mental tokens have the same content, then they have the same form. Since Fodor has stipulated that contents differ only if they can be identified with different forms, evidently he subscribes to thesis (i), whose contrapositive maintains if mental tokens do not have the same content, then they do not have the same form. Although we have observed that Fodor does not elaborate how distinct thoughts are connected to distinct tokens, we shall assume that he intends that, say, similar surface grammars sometimes conceal differences in meaning that are disclosed by an investigation of their deep structures, which means that their parse trees, if they were parsed, would differ, etc.

Indeed, there seem to be at least three different ways in which tokens having the same form might be exhibited as possessing different content, each of which appears to be successively less and less syntactical in nature:

- (1) *the parsing criterion*, according to which tokens have the same content only if they have the same parse trees;
- (2) *the substitutional criterion*, according to which tokens have the same content only if they are mutually derivable by the substitution of definiens for definiendum; and,
- (3) *the functional role criterion*, according to which tokens have the same content only if they fulfill the same causal role in their effects upon behavior relative to all possible situations.

Each of these, no doubt, requires unpacking to be clearly understood. Since they accent syntactical, semantical, and behavioral dimensions of language and mentality, in turn, we shall consider them separately in the following, beginning with the parsing criterion in relation to the syntactical approach.

Distinctive versions of the computational conception might defend one of these conditionals but abandon the other, while retaining their computational character; indeed, that appears to be Fodor's own position here. But it is difficult to image how a position that abandoned *both* of these theses could still qualify as a 'computational' conception. Thus, in order to display the poverty of the computational approach, arguments will be advanced to show that both of these conditionals are *false*. Moreover, the reasons that they are false are surprisingly obvious, but none the less telling. The first of these conditionals falls prey to the problem of ambiguity, while the second succumbs to the problem of synonymy, as numerous examples display.

The first conditional claims that mental tokens have the same content if they have the same form. This thesis would be false if it were ever the case that some mental tokens have the same form, yet differ in their content. If ordinary sentences in the English language qualify as 'mental tokens', therefore, then examples involving ambiguous words – such as 'hot', 'fast', etc. – generate counterexamples whose status as ambiguous sentences is not likely to be challenged. Consider the following specific cases as illustrations:

Example (A): Imagine a very warm summer afternoon, as a group of guys are conversing about a shiny red convertible. One casually remarks,

(a) 'That car is hot!'

Clearly, at least two meanings (contents) might be intended by remark (a), since it might be interpreted as a comment on the temperature of the car,

(b) 'That car has heated to over 100 degrees Fahrenheit!'

but might be meant to convey its status as a recent and illegal acquisition:

(c) 'The cops are out looking for that car everywhere!'

In this case, tokens of the same form, (a), could thus have different content.

Example (B): Sitting on the steps of the stadium, two girls are engaged in animated conversation. As a young man passes by, one says to the other,

(d) 'John is fast!'

Once again, at least two meanings (contents) might be intended by remark (d), since the speaker might be commenting on John's prowess as a sprinter,

(e) 'John can beat almost everyone else at the 100 yard dash!'
but it might also be intended as a characterization of his dating behavior,

(f) 'John wants to go farther faster than anyone I have dated!'

Once again, therefore, tokens of the same form (d) could differ in meaning.

Initially, ambiguous tokens, such as these, might seem to pose no problem for the computational conception, since the parsing criterion could be employed to discriminate between them. Since the construction of a parse tree presumes the availability of a suitable grammar for this purpose [cf. Winograd (1983)], let us adopt the following very elementary grammar, G :

Grammar G:

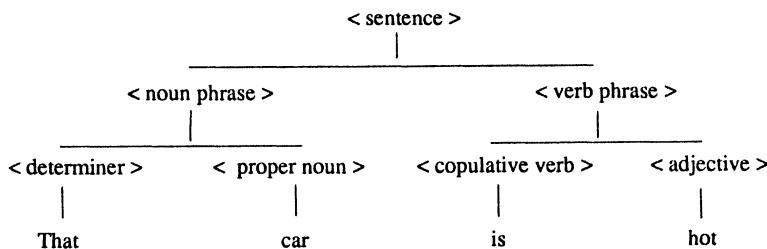
```

<sentence> → <noun phrase><verb phrase>
<noun phrase> → <proper noun>
<noun phrase> → <determiner><proper noun>
<verb phrase> → <copulative verb><adjective>

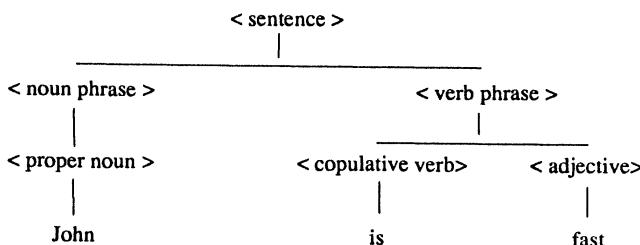
```

which, for this exercise, is all of the grammar that happens to be required.

It is important to observe, therefore, that the parse tree for example (a) turns out to have the following structure,



while the parse tree for example (d) has the following different structure,



Thus, in terms of their parse trees, example (a) and example (d) clearly differ in their parse trees and thus differ in their form, which presumably provides *prima facie* evidence in support of the computational conception. But this is the case only if they also differ in their content. And we know that they differ in their content, because we have assumed they are in English!

The point is that, while it is indeed true that, relative to grammar G , the parse trees for sentences (a) and (d) are indeed different, that supports thesis (i), which asserts that tokens with different content have different forms, *only if* sentences (a) and (d) do have different content! We have taken for granted that the sentences, 'That car is hot!' and 'John is fast!', have different meanings and therefore ought to have different forms – provided thesis (i) is true. If, however, the object referred to by the noun phrases, 'that car' and 'John', were the same and the properties ascribed by the verb phrases, 'is hot' and 'is fast', were the same, then the results attained above would count as a counterexample *against* thesis (i) rather than as evidence *for* it! The effect obviously depends upon and varies with the relevant language.

Let us assume that all of these examples are well-founded, so there is no basis for denying that, in ordinary English, sentence (a) is often used to mean what is expressed by sentence (b) but also to mean what is expressed by sentence (c) – and, analogously, for sentence (d) in relation to sentences (e) and (f). Then these examples clearly establish that thesis (ii) is false, insofar as there are at least some sentences in English that have the same content but different forms, which would be impossible if mental tokens having the same content have to have the same form. It is evidently not the case that, if mental tokens differ in their form, then they differ in their content. This thesis of computationalism cannot cope with the problem of synonymy.

Now imagine an impoverished version of English, sub-English-4, say, in which the *only* words in the vocabulary happen to be those that are found in sentence (a) – or, alternatively, sub-English-3, those in sentence (d). Naturally, these are fanciful scenarios, since, for a variety of reasons, these are not likely to really be the *only* words in anyone's vocabulary. Nevertheless, they represent a class of cases of the relevant kind. Under these circumstances, would it not be possible that (a) could sometimes be used to mean (b) and sometimes to mean (c) or that (d) could sometimes be used to mean (e) and sometimes to mean (f)? Of course, it would not be possible to *articulate* the contemplated difference: would that mean *no* such difference could exist?

There are strong reasons for thinking it would not! Consider, after all, that some of the most frequently used exclamations in the English language vary

in their meaning with the context of their use: 'Damn!', for example, is sometimes used as an expression of distress (discomfort or pain), but it also occurs as an expression of relief (happiness or joy). Here the same word is being used with very different content. Similarly, other criteria, such as the functional role criterion, could differentiate between different meanings for sentence (a): if the gang were to cease talking and scatter when a police car approached, that would tend to indicate that (a) was being used in sense (b); if they cracked an egg on the hood of the car when they wanted to eat it for a snack, that would tend to indicate that (a) was being used in sense (c); etc.

In cases of this kind, these differences in meaning (or content) would indeed exist, even though they could not be articulated *within those languages*. If these examples are acceptable, therefore, then it seems clear that thesis (i) is also false, insofar as at least some sentences in English (sub-English-4 or sub-English-3) have the same form but different contents, which would be impossible if mental tokens having the same form have to have the same content. This thesis of computationalism cannot cope with the problem of ambiguity. Appeals to dependence of meaning upon context, moreover, can be extended to larger classes of sentences through considering successively larger and larger units of discourse. But it should not be overlooked that successively larger and larger units of discourse remain vulnerable to problems of ambiguity, to problems of synonymy, and to problems of context.

There are no grounds at all to believe that longer and longer sentences (sequences of symbols, strings of marks) are necessarily or inevitably less and less ambiguous, as the computational conception would suggest. Sets of sentences constituting paragraphs and sets of paragraphs comprising documents (such as the *Constitution*) remain matters of great debate – even when attempts are made to understand them within their historical context! It would therefore be a mistake to imagine that appeals to 'context' would be more likely to support the computational conception than to undermine it. If these reflections are well-founded, therefore, then the computational conception does not appear to have a great deal to recommend it; for the twin theses that define that position are relatively clearly and decisively flawed.

2. THE REPRESENTATIONAL CONCEPTION: SEMANTICAL MODELS OF THE MIND

Hilary Putnam (1975) differentiates between psychological (or 'mental') states for which the ascription of that state (or 'token') does *not* presuppose

“the existence of any individual other than the subject to whom that state is ascribed” (psychological states in *the narrow sense*) and the rest (psychological states in *the broad sense*) [Putnam (1975), p. 136]. As Fodor remarks, “Narrow psychological states are those individuated in light of the formality condition; viz. without reference to such semantic properties as truth and reference” [Fodor (1980), p. 331]. In view of the arguments presented in Section 1, it would not be unreasonable to wonder whether any account of the nature of language and of mind which, like computational accounts, restricts itself to psychological states in the narrow sense could possibly be adequate. In order to appreciate its appeal, therefore, it is crucial to consider precisely how a purely syntactical account can accommodate content.

Recall that computational conceptions not only require some strong relationship between the form of a token and its content but also need to explain how *any* specific content comes to be identified with *any* specific form. There appear to be at least three possible solutions to this problem, namely:

- (4) *by nature*, according to which the connection between tokens of a specific form and their specific content is a function of the laws of nature;
- (5) *by convention*, according to which the connection between tokens of a specific form and their specific content is a function of the practices, customs and traditions of a social group; and,
- (6) *by habituation*, according to which the connection between tokens of a specific form and their specific content is a function of the habits, skills and dispositions of individual token-users.

Indeed, the approach among these three that blends most harmoniously with syntactical models of the mind is the thesis that these relations are ‘natural’.

Once mental tokens have been ascribed content by one or another of the modes indicated above, however, it becomes increasingly difficult to distinguish ‘computational’ from ‘representational’ conceptions. Stich, for example, has sought to separate ‘strong’ from ‘weak’ representational accounts:

Unlike the STM, however, the weak RTM [Representational Theory of the Mind] insists that these syntactic objects *must have content or semantic properties* ... The weak version claims only that every token mental state to which a cognitive theory applies has *some* content or *some* truth condition... The stronger version of the doctrine agrees with the weaker version in requiring all mental state tokens to have

content or truth conditions. But it goes on to claim that these semantic features are correlated with the syntactic type of the token. [Stich (1983), p. 186].

Thus, (4), (5) and (6) reflect successively weaker and weaker modes of correlation between the syntactical type of a token and its semantical content.

A purely syntactical conception of language and mentality, presumably, would be an object of limited interest, since it would be unable to accommodate the ideas most fundamental to cognition and communication: content, information and belief. Indeed, even Haugeland (1981), who has championed the computational conception of minds as 'automated formal systems', has acknowledged the necessity for the formal tokens of those systems to possess semantic content: "Given an appropriate formal system *and interpretation*, the semantics takes care of itself" [Haugeland (1981), p. 24]. The problem that remains, therefore, is to explain how these correlations occur, without which the syntactical approach would simply fail to distinguish 'interpreted' from 'uninterpreted' formal systems [Hempel (1949a), (1949b)].

Fodor (1975) attempts to resolve this problem at a single stroke by introducing the notion of *the language of thought*, envisioned as an unlearned language that is both innate and species-specific. Among the most important theses upon which his position depends are (i) that learning a language presupposes the possession of a prior language; (ii) that learning a language cannot be a matter of acquiring dispositions; and (iii) that this innate language functions as a meta-language for learning other languages [Fodor (1975), pp. 64–79]. The key to Fodor's position is the assumption that learning a language requires learning the truth conditions for the sentences that can occur in that language: "... learning (a language) L involves learning that Px is true if and only if x is G for all substitution instances. But notice that learning that could be learning P (learning what P means) only for an organism that already understands G " [Fodor (1975), p. 80]. Having made this assumption, Fodor finds himself confronted by the unpalatable choice between the existence of an infinite hierarchy of meta-languages (for each meta-language in turn) and the existence of an unlearned language that serves as a base case.

Given exclusive commitment to truth condition semantics in the sense adumbrated here, Fodor suggests that learning a language is "literally a matter of making and confirming hypotheses about the truth conditions associated with its predicates" [Fodor (1975), p. 80], supporting the following claim:

Either it is false that learning L is learning its truth definition, or it is false that learning a truth definition for L involves projecting and confirming hypotheses about

the truth conditions upon the predicates of L , or no one learns L unless he already knows some language different from L but rich enough to express the extensions of the predicates of L . [Fodor (1975), p. 82]

Notice that it is the commitment to truth-functional semantics as an exclusive access to the meaning of a sentence in L combined with the theoretical necessity to block an infinite regress of meta-languages for meta-languages that anchors Fodor's position. No one, presumably, would want to deny that learning a language involves projecting and confirming hypotheses about the predicates of L ; the question is, does this have to be done as Fodor suggests?

On its face, Fodor's position appears to be very difficult to swallow. Its general character parallels Plato's theory of knowledge as recollection, which posits an Eternal Mind in which mortal minds participate before birth. Since the Eternal Mind is the respository of all knowledge and mortal minds participate in the Eternal Mind, all knowledge resides in every mortal mind before birth. Aware that not all mortal minds appear to be all-knowing, Plato contends that the trauma of birth induces forgetfulness in each of us, but that different experiences in life trigger off 'recollections' of that lost knowledge. An alternative to Plato's account, of course, would be the theory that knowledge is acquired through experience, so that the very 'experiences' in life that (for Plato) trigger off 'recollections' could be the mechanisms through which knowledge is actually acquired, an account that is far more elegant.

In an analogous fashion, Fodor posits a universal language of thought of which every (neurologically normal) human being is a possessor. In order to learn any ordinary language, it is necessary to discover (through experience) the truth conditions for each of the predicates in that language within the language of thought. The language of thought itself, moreover, must be infinitely rich and extraordinarily complex, since it must have the capacity to encompass each new predicate as it is introduced into its previously-im-poverished ordinary language successor (in response to the discovery and growth of science and technology, for example). Since not all speakers of ordinary language display similar linguistic ability, evidently they simply have not exercised equal ingenuity in discovering the truth conditions that lie undiscovered in their mental language as an as-yet-unrealized resource.

Since experience and ingenuity as well as the language of thought are necessary conditions for learning an ordinary language on Fodor's account, an alternative would be the theory that languages can be learned through experience and ingenuity, so that the very 'experiences' in life that lead to

projecting and confirming hypotheses about the predicates of L might occur even in the absence of a language of thought! For this to be possible, however, Fodor's argument must have at least one false premise; and, indeed, a little reflection suggests that Fodor may have begged the question in moving from the premise that learning P (learning what P means) can only occur for 'an organism that already understands G ' to the conclusion that learning P (learning what P means) requires learning meta-linguistic truth conditions for P . For that conclusion would follow *only if* there were no other way for an organism to 'understand' the G -phenomenon that P happens to describe.

An alternative to Fodor's theory of learning a language, in other words, would be to take seriously that learning presupposes some sort of 'understanding', without assuming that the form of understanding involved here has to be *linguistic*. As infants and children, we frequently – even typically – learn to do things (suck a nipple, bounce a ball, smile a lot) without having any name or label for the habits, skills or activities thereby performed. It should not be especially surprising, therefore, that when (initially unfamiliar) words are associated with (already familiar) things, including patterns of behavior that we happen to have displayed, it does not demand extraordinary ingenuity or vast experience for a (neurologically normal) human being to learn forms of language that are appropriate to their age and past experience. The problem for Fodor is the same as the problem for Plato.

The same conclusion follows from a different starting point. One of the virtues of computational and of representational accounts would appear to be their capacity to exploit (what we shall refer to as) 'the inferential network' model of language and mentality. As Christopher Maloney observes,

Computationalism considers the mind to be an inferential system. That clearly presupposes the existence of structures over which the inferences are defined ... [and] the only things that seem physically fit to function as elements in material inferences are sentences. [Maloney (1988), p. 56]

Indeed, the meaning of a word or a sentence can be related to its place within a network of syntactical and of semantical relations, a paradigm of which is the standard conception of a *definition* as an entity consisting of two parts, a word, phrase or expression to be defined ('the definiendum') and the word phrase or expression by means of which it is defined ('its definiens'), where both the definiendum and the definiens are envisioned as *linguistic* entities.

The catch, of course, is that it is impossible for every word in a language to

be defined on pain of either an infinite regress or definitional circularity, as everyone would acknowledge. Yet every defined term can be replaced in principle by some sequence of undefined terms with which it is synonymous. Thus, by the substitutional criterion, different mental tokens have the same content when they are mutually derivable by the exchange of *definiens* for *definiendum*. Even representational accounts, however, afford no solution to the problems of ambiguity and context that we have considered above.

Assuming that words like 'fast,' for example, are defined terms within the corresponding language, sentences like (d) and (e) or (d) and (f) could well be mutually derivable in harmony with the substitution criterion and presumably would possess the same content in accord with that standard. But even sentences like (f) remain ambiguous, since precisely 'where' John 'wants to go' might have social, sexual, or professional connotations, among others, depending upon context. The relations that obtain between the different elements that collectively comprise an 'inferential network' do not fix their content. The problem that therefore remains is accounting for the meaning of the undefined (or 'primitive') terms of that language. By positing a 'language of thought', of course, Fodor attempts to finesse this difficulty, which would resurrect itself *within the language of thought* were it not for the thesis that the meaning of these tokens has been fixed 'by nature'!

The question sounds funny on its face; yet I want to contend that this is the issue that poses the deepest concern for both computational and representational conceptions. The answer, after all, apparently consists of those habits, skills and practices by virtue of which the words we use are related to the world around us: we seldom think about defining words like 'wood', 'hammer', and 'nail', because we can use them – and can know how to use them – without the intervention of other linguistic forms. With respect to the primitive terms that occur in any ordinary language, we must discover how they are used rather than ask for their linguistic meaning, precisely as Wittgenstein proposed. From this point of view, therefore, knowledge of language is far more adequately envisioned as a skill than as a state (as a matter of 'knowing how' rather than of 'knowing that'), which itself may be the fundamental misconception lying at the foundation of these accounts.

A theory of language and mentality that could handle only uninterpreted formal systems, no doubt, would completely fail to satisfy the most elementary desiderata for theories of that kind. Yet it is important to notice a difference between at least two ways in which content could accompany a syntactical conception 'by nature'. A theory of the first type might assert the existence of laws of nature relating the sentential tokens that happen to occur

in ordinary languages, such as English, German and French, to mental states with specific content. The very existence of ordinary languages with such very different grammars and vocabularies, however, suggests that an approach of this kind cannot possibly be correct. There do not appear to be any grounds for accepting any semantic theory with these general features.

A theory of the second type, however, might assert the existence of laws of nature relating mental states to specific content, while making connections between those mental states and the sentential tokens that occur in ordinary languages a matter of convention. An account of this kind, in effect, would posit a set of *linguistic primitives* for an ordinary language (consisting of the undefined words in its vocabulary) and a set of *psychological primitives* for a corresponding mental language (consisting of the tokens that are related to their content 'by nature') [cf. Fodor (1975), p. 124]. Obviously, Fodor's own account is of this general kind. A theory of yet a third type, however, might deny the existence of laws of nature of either kind, asserting instead that all of these connections are established either by conventions or else by habits. Only a theory of this kind offers the promise of resolving the difficulties that have been discussed above. But such an account, which emphasizes the role of individuals and of communities in establishing the significance of signs (as their users), could not appropriately qualify as a representational conception. Perhaps a pragmatical conception would provide a more promising approach.

3. THE DISPOSITIONAL CONCEPTION: PRAGMATICAL MODELS OF THE MIND

If these arguments are well-founded, then the dispositional conception appears to be right-headed on the whole, insofar as it provides an account that is completely compatible with the identification of the meaning of the primitives of a language with the linguistic habits that effect these connections between language and the world. Fodor's defense undoubtedly would be a thesis that we have merely mentioned in passing. For Fodor contends that learning a language cannot be a matter of acquiring dispositions: "If anything is clear", he maintains, "it is that understanding a word (predicate, sentence, language) isn't a matter of how one behaves or how one is disposed to behave" [Fodor (1975), p. 63]. And I would freely admit that some of the classic positions on the nature of dispositions, such as those of Ryle (1949), of Skinner (1957) and of Quine (1960), are unable to cope with his criticism.

These accounts suffer from a variety of maladies, which arise because the conceptions they present are behavioristic, reductionistic and extensional in kind. The nature of dispositions as they are understood here, by comparison,

is non-behavioristic, non-reductionistic and non-extensional [Fetzer (1981), (1986)]. Nevertheless, the adequacy of the analysis that follows can be appraised in relation to Fodor's principal argument against dispositions:

Behavior, and behavioral dispositions, are determined by the interactions of a variety of psychological variables (what one believes, what one wants, what one remembers, what one is attending to, etc.). Hence, in general, any behavior whatever is compatible with understanding, or failing to understand, any predicate whatever. Pay me enough and I will stand on my head if you say 'Chair'. But I know what 'is a chair' means all the same. [Fodor (1975), p. 63]

For no dispositional conception of language and mentality should be adopted unless it can explain what is right and what is wrong with this position.

All of this would be so much 'smoke and mirrors', however, were it impossible to demonstrate the benefits that accrue from adopting a pragmatic point of view. Perhaps its most crucial features were recognized by Peirce, who accentuated the place of beliefs as causal elements affecting behavior.

Our beliefs guide our desires and shape our actions... The feeling of believing is a more or less sure indication of there being established in our nature some habit which will determine our actions... Belief does not make us act at once, but puts us into such a condition that we shall behave in some certain way, when the occasion arises. [Buchler (1955), pp. 9-10]

An analysis of this kind clearly falls within the broad tradition of functional conceptions of meaning and of mind, where the difference between this account and others of this general type is the specific role assigned to dispositions in understanding the nature of acts, in general, and of speech acts, in particular. While dispositional conceptions are functional accounts, in other words, not all functional accounts are dispositional conceptions – nor, indeed, do other dispositional accounts possess the special characteristics of this one.

Peirce defined a 'sign' as a something that stands for something (else) in some respect or other, where there are fundamental differences in the ways in which something can 'stand for' something (else). Thus, in particular, he distinguished between three different classes or varieties of signs as follows:

- (7) *icons*, which are signs that stand for other things by virtue of a relation of resemblance between those signs and that for which they stand;
- (8) *indices*, which are signs that stand for other things by virtue of being either causes or effects of those things for which they stand; and,

- (9) *symbols*, which are signs that stand for other things by virtue of conventional agreements or habitual relations between those signs and that for which they stand.

Photographs, paintings and statues, for example, are icons, while smoke and fires, symptoms and diseases, etc., are instances of indices. The words that make up ordinary languages, such as 'dog' and 'chair', by contrast, neither resemble nor are causes or effects of that for which, as symbols, they stand.

Another reason for disputing the thesis that thought requires language (in the sense that ordinary languages are languages), therefore, is that their vocabularies stand for that for which they stand because of conventional or habitual connections between those words and those things for which they stand, that is, they exemplify only *one* among at least three different types of ways that something can stand for something else. This, in turn, raises the intriguing possibility that there might be more than one type of mentality corresponding to more than one type of system of signs, where *semiotic systems* of Type I utilize icons, of Type II utilize icons and indices, and of Type III utilize icons, indices and symbols [Fetzer (1988a), (1988b)]. Moreover, there are grounds for thinking there may be grades of mentality that are higher than that of semiotic systems of Type III [Popper (1978), (1982)].

Nevertheless, the theory of signs provides a foundation for the theory of belief only if (what we shall call) 'a theory of cognition' ties them together. Thus, while signs provide modes of reference for objects and for properties, they lack the assertive character of beliefs, sentences, and propositions, i.e., they are neither true nor false. The connection between signs and beliefs for a semiotic system, therefore, appears to be a causal process that arises when a system becomes conscious of a sign in relation to its *other internal states*, including its pre-existing motives and beliefs. When such a system becomes conscious of something that functions as a sign for that system, its cognitive significance results from causal interaction between that sign and these internal states, which constitute its *context*. From this point of view, therefore, the content (or meaning) of a mental state (or token) cannot be fixed independently of consideration of the context provided by a semiotic system, apart from which even its constituent signs possess no significance.²

Since the term 'concept' would be useful to refer to the meaning of any mental token (whether sentence, sign, or belief), let us adopt it here. Thus, a complete account of the content of a concept for a specific semiotic system (if it were possible) would be provided by an inventory of all of the kinds of

behavior toward which that system would be disposed under all of the different kinds of contexts within it might find itself. The conception of 'behavior' required by this construction, however, must be broad enough to encompass mental effects among its manifestations, which occur, for example, when someone 'changes their mind' [cf. Fetzer (1988b), p. 139]. When the content of a concept would be displayed in various ways under infinitely varied conditions, then any merely finite description of its significance could never be more than partial and incomplete. The most suitable approach toward understanding the content of a concept is in relation to its causal role:

(Thus,) the most perfect account of a concept that words can convey will be a description of the habit which that concept is calculated to produce. But how otherwise can habit be described than by a description of the kind of action to which it gives rise, with a specification of the conditions and of the motive? [Buchler (1955), p. 286]

Indeed, it seems clear that, in the case of human semiotic systems, the range of behavioral manifestations that a concept has for a system would have to vary across (the complete sets of) motives, beliefs, ethics, abilities, capabilities and opportunities that influence its behavior as a complex causal system.

From this perspective, the theory of meaning presupposes the theory of action. The theory of action that shall be adopted here takes it for granted that *human behavior* and *human actions* are not identifiable, insofar as the class of human actions is restricted to the class of human behaviors that are brought about – possibly probabilistically – by the causal interaction of one's own motives, beliefs, ethics, abilities and capabilities, where the success (or failure) of those efforts tends to depend upon and vary with the opportunities with which we are confronted (including, in particular, with whether or not the world is as we believe it to be, i.e., as a function of their truth). Ultimately, of course, it may be important to distinguish more precisely between these assorted types of factors, but let us assume they form a complete set.

To illustrate the character of the account that I am endorsing, observe that a marksman who wants to hit his target and believes that his target is present and who does not rule out firing at this target on moral grounds can hit his target only when his skills are equal to the task, his rifle and ammunition are available and the target itself is within his vicinity [Fetzer (1986), p. 106]. When an individual happens to be neurologically impaired, physically

restrained, morally debauched, deliberately misinformed, etc., then the kinds of behavior that they tend to display under otherwise similar conditions varies from those that tend to be displayed by individuals who are not neurologically impaired, physically restrained, morally debauched, etc. The behavior someone displays on a specific occasion thus results from the complete set of relevant factors present on that occasion, where a factor is relevant if its presence or its absence on that occasion made a difference to the strength of the tendency for that person to display behavior of that kind.

Within the scope of this conception, the content (or the meaning) of a specific sign (or token) can be captured by identifying its causal role in influencing different kinds of behavior under different kinds of conditions, where those conditions (for human systems) tend to be complex. Thus, a few examples may serve to illustrate the conception recommended here:

Example (C): John has the misfortune to be confronted by a burglar in his own apartment. The burglar, whom John does not recognize, menaces him with a knife, so he escapes by climbing out the window, which is open.

Example (D): Mary has the misfortune to be confronted by a burglar in her own apartment. The burglar, whom Mary does not recognize, menaces her with a knife, but she has a broken leg and cannot climb out the window.

In appraising whether or not the same sign (the burglar with the knife) has the same content (or meaning) for John and for Mary, a comparison has to be drawn, not between their *actual* behavior (which was obviously different) but between their *dispositions toward behavior* (how they would have behaved, if they had the chance, relative to the same motives, beliefs, etc.), which requires intensional (subjunctive and counterfactual) rather than extensional (historical and indicative) formulations. If John and Mary would have behaved in the same ways across every complete set of relevant conditions when conscious of that sign (when these contexts were the same for both), then this sign would possess the same content (or meaning) for both, even though the behavior they actually displayed may have been different! In particular, if Mary would have escaped out the open window, if she had not had a broken leg; if John would have done whatever Mary actually did, had he found himself in her situation (including her abilities and capabilities); etc., then that sign would have had the same content for both of them.

What this implies is that Fodor is mistaken in thinking that there could be *no* behavioral constancy that would provide a foundation for a dispositional

account. Fodor's mistake results from adopting an excessively behavioristic and extensional account of dispositions. "Pay me enough and I will stand on my head if you say, 'Chair,'" says he. "But I know what 'is a chair' means all the same." Exactly! So too for everyone else whose motives and morals would inspire them to similar behavior – knowing what 'is a chair' means just as well! It also suggests the solution to a difficulty observed by Donald Davidson long ago, according to which the meaning of a token (sign) cannot be a disposition to use specific words in specific ways, *simplicitur*, if only because not everyone speaks English! Of course, he is right, except that our dispositional conception only requires that different speakers use similar words on similar occasions in similar contexts – *provided they have the same linguistic abilities!* Otherwise, no such similarity needs to follow.

Hence, if we want to isolate special kinds of causal factors, such as the content (or meaning) of specific beliefs B_1, B_2, \dots , then we can do so by holding constant those other beliefs B_m, B_n, \dots , motives, M_1, M_2, \dots , ethics E_1, E_2, \dots , abilities A_1, A_2, \dots , capabilities C_1, C_2, \dots , and opportunities O_1, O_2, \dots , whose presence or absence makes a difference to the (internal or external) behavior that would be displayed by that system, given the presence of B_i (where i ranges over 1,2,...). Then the content (or meaning) of a specific belief B_2 , say, is the totality of tendencies that the system would possess in the presence of that belief, and the difference that having that belief rather than some other, say, B_1 , is the difference between the totality of tendencies that that system would possess in the presence of B_1 and the totality of tendencies that that system would possess in the presence of B_2 !

For any specific semiotic system, therefore, a sign, S , stands for something x for that system rather than for something else y if, and only if, the strength of the tendencies for that system to manifest behavior of some specific kind when conscious of S – no matter whether publicly displayed or not – differs in at least one context; otherwise, there is no difference between x and y for that system. When two signs, S_1 and S_2 , possess exactly the same meaning (as two tokens of exactly the same type), then the strength of the tendencies for that system to manifest behavior of various kinds when conscious of S_1 must be the same in every context as is the strength of its same tendencies when conscious of S_2 . This account therefore allows – indeed, it requires – that 'sameness of meaning' be amenable to *degrees of similarity* that occur when two tokens are tokens of some of the same types but not of all. When comparisons of (what might be called) 'cognitive significance' in *its narrow sense* (encompassing differences in meaning other than purely linguistic ones, involving the specific words used, their precise sounds, etc.) are

desired, that measure can be obtained, in principle, by subtracting all these purely linguistic behavioral phenomena from (what might be called) 'cognitive significance' in *its broad sense* (which encompasses all meaning). Indeed, these conceptions afford a foundation for resolving issues of ambiguity and of synonymy across the board within a dispositional framework.

Methodologically, at last, this conception exemplifies what Carl G. Hempel describes as, "the epistemic interdependence of belief and goal attributions" [Hempel (1962), Sec. 3.3] – or, better, "the epistemic interdependence of motive, belief, ethics, ability, capability and opportunity ascriptions". This means that, in order to subject an hypothesis about causal factors of any of these kinds to empirical test, it is necessary to make assumptions concerning the simultaneous values of each of the others. Fortunately, this result, which is not theoretically avoidable, is not theoretically objectionable. But if such a conception is even roughly right-headed in its approach, it provides a rather striking explanation for the inherent complexity of social and psychological phenomena. From this point of view, the complexity of the phenomena with which social science must contend in comparison to that with which natural science must contend becomes strikingly apparent – a remarkable outcome!

4. LANGUAGE AND MENTALITY: CONCLUDING REFLECTIONS

During the course of this investigation, we have undertaken an exploration of the solution space for the problem of discovering an adequate conception of the nature of language and mentality. Attention has been given to approaches of three distinct types, computational, representational, and dispositional conceptions, which accent syntactical, semantical and pragmatical aspects of the phenomena with which they are concerned, respectively. We have discovered that the computational conception, even when complemented by the parsing criterion, adopts an extremely implausible theory of the relation of form to content, which cannot contend with problems of ambiguity, with problems of synonymy, or with problems of context. We have also discovered that the representational conception, which benefits from an 'inferential network' approach and from the substitutional criterion, cannot resolve the underlying difficulty of fixing the meaning of primitive language.

These reflections have led to an exploration of the dispositional conception that, unlike its alternatives, appears to succeed where they have failed. In particular, by appreciating the role of habits, skills and dispositions in

establishing connections between language and the world, the pragmatical approach has the capacity to deploy the functional role criterion in dispatching problems of ambiguity and of synonymy, while taking proper account of the place of context. It overcomes the *prima facie* case against dispositional conceptions by embracing an account of dispositions that is at once non-behavioristic, non-reductionistic and non-extensional, which enables it to overcome not only Fodor's complaints but Davidson's objection as well. Thus, an analysis of this kind, which unpacks the content (or meaning) of mental tokens by means of their causal role within an intensional framework, is not vulnerable to the criticisms that undermine different accounts, including Quine (1975).³

One of the most important consequences that attend this investigation, moreover, emanates from the crucial role of individual sign-users as semiotic systems. For the notion of an *idiolect* as the sign system utilized by a single sign-user turns out to be more fundamental theoretically than is that of a *dialect* (understood as a regional phenomenon) or that of a *language* (if understood as a social group phenomenon). Thus, there is nothing here that inhibits the prospects for the existence of 'private languages', in the sense of constellations of dispositions for speech and other behavior that reflect surface manifestations of possibly unique correlations of primitives to the world. Indeed, the results that have been uncovered here clearly suggest that convention has a secondary role to play by contrast with habituation, promoting communication by providing a social mechanism for resolving differences and for codifying practices concerning what does and does not qualify as 'standard usage' within particular language-using communities.

Another important consequence attends the realization that physical systems can be distinguished as systems of (shall we say) 'fundamentally different' kinds when the specific type of factors that make a difference to the behavior of those systems varies from case to case. Human beings, from this perspective, qualify as motive, belief, ethics, ability, capability and opportunity types of systems, since these reflect the range of causal factors that affect the behavior of systems of this kind. Digital computers, by comparison, seem to be electronic, magnetic, hardware, firmware, software and input types of systems, since these reflect the range of causal factors that affect the behavior of systems of that kind. The proper measure of similarity and difference, once again, of course, has to be subjunctive and counterfactual rather than historical and descriptive in relation to complete sets of relevant conditions that influence systems of these kinds.

Just as human beings and digital machines qualify as different types of

physical systems, so too may they qualify as different sorts of semiotic systems. Indeed, if systems qualify as semiotic by virtue of establishing connections between the tokens they employ and the potential behavior they might display, then the semiotic abilities of humans and computers appear to be distinct. Human beings are fundamentally similar, from this point of view, when they are capable of acquiring and of manifesting all and only the same semiotic tendencies under the same causal conditions. Similarly, digital machines are fundamentally similar, from this point of view, when they are capable of performing all and only the same computations (calculations and operations) under the same causal conditions – which obviously includes the programs that they have the ability to run.

All fundamentally similar human beings and digital machines, however, need not actually acquire and manifest all and only the very same semiotic and computational tendencies, unless they are under the same causal conditions at all times – and, even then, not unless their predispositions to acquire and their tendencies to manifest those tendencies are deterministic rather than probabilistic. Thus, digital machines would be ‘fundamentally similar’ to human beings only if they were subject to the same range of factors, which is plainly not the case. Not the least of the benefits that follow from adopting the dispositional point of view, therefore, is that it provides an explanation for the deeply held intuition that human beings and digital machines really are ‘fundamentally different’ as types of knowledge, information and data processing causal systems.

While most students of artificial intelligence tend to believe that AI concerns how human beings do think rather than how they should think, the results of this investigation suggest another possibility. If, after all, human beings and digital machines *are* ‘fundamentally different’ as we have discovered above, then what grounds remain in support of the view that digital machines and human beings *can* process knowledge, information or data in similar ways? The great debate between the ‘strong’ and the ‘weak’ conceptions of AI, it appears, may rest upon a misconception, which would indeed be the case if systems of one kind (digital machines) are incapable of functioning in the same way as systems of the other kind (human beings). For, if this were the case, then it might not only be false that AI concerns how we do think but also false that AI concerns how we should think. If this were the case, then the proper conclusion to draw might be that AI strives to develop machines that can solve problems in ways that are not accessible to human beings, in which case it might be maintained that the aim of AI is the creation of new species of mentality.

NOTES

* This paper originated as an informal lecture entitled 'Cognitive Science and Epistemic Inquiry', which was delivered at the University of Georgia as the Keynote Address for the First Annual Southeastern Graduate Students' Philosophy Conference held on 11-12 April 1986. Special thanks to everyone responsible for the organization of that meeting, especially Terry L. Rankin. I have benefitted from the stimulating remarks of two anonymous referees and added some footnotes thanks to Charles E.M. Dunlop and to David Cole.

¹ One caveat. The argument might be made that the objections lodged in Section 1 do not affect the language of thought, precisely because in that language (or 'mentalese') our *thoughts* are always unambiguous. Even if this claim were correct, it would depend upon the nature of thought as a computational, representational or dispositional conception might unpack it. Fodor, however, suggests that the language of thought is 'very similar' to ordinary languages in all the relevant respects [Fodor (1975), p. 156].

This not only justifies the use of sentences from an ordinary language like English to exemplify the difficulties confronting his position but also hints that, if there *is* some unambiguous level of understanding, it is far more likely to be dispositional than to be like anything that Fodor has in mind. Only a pragmatical conception of language and mentality provides a theoretical foundation that might possibly explain why thought should seem to be so very different from the signs that we employ to express it.

² The use of the term 'conscious' may need some explanation here, since I do not thereby intend to preclude the potential influence of preconscious or of unconscious mental states. A sign-using system is *conscious* (with respect to signs of a certain kind) when it has both the ability to utilize signs of that kind and the capability to exercise that ability, where the presence of signs of that kind within the appropriate causal proximity would lead – invariably or probabilistically – to an occurrence of cognition [Fetzer (1990)].

³ One of the principal motives for pursuing an analytical problem-space approach for this investigation is that some of the most influential figures working within this field have gradually evolved in their positions, Fodor especially. Fodor (1987), for example, yields a functional analysis within common-sense psychology; yet he persists in maintaining that "there has to be a *language of thought*" [Fodor (1987), pp. 135-154]. On this point, I believe he is completely mistaken. The only presupposition for learning a language appears to be species-specific *predispositions* [cf. Fetzer (1985)].

REFERENCES

- Buchler, J.: 1955, *Philosophical Writings of Peirce*, New York: Dover Publications.
- Carnap, R.: 1939, *Foundations of Logic and Mathematics*, Chicago, IL: University of Chicago Press.
- Charniak, E. and McDermott, D.: 1985, *Introduction to Artificial Intelligence*, Reading, MA: Addison-Wesley.

- Fetzer, J.H.: 1981, *Scientific Knowledge*, Dordrecht, Holland: D. Reidel.
- Fetzer, J.H.: 1985, 'Science and Sociobiology', in Fetzer, J. (ed.), *Sociobiology and Epistemology*, Dordrecht, Holland: D. Reidel, pp. 217–246.
- Fetzer, J.H.: 1986, 'Methodological Individualism: Singular Causal Systems and Their Population Manifestations', *Synthese* 68, pp. 99–128.
- Fetzer, J.H.: 1988a, 'Mentality and Creativity', *Journal of Social and Biological Structures* 11, pp. 82–85.
- Fetzer, J.H.: 1988b, 'Signs and Minds: An Introduction to the Theory of Semiotic Systems', in Fetzer, J. (ed.), *Aspects of Artificial Intelligence*, Dordrecht, The Netherlands: Kluwer Academic Publishers, pp. 133–161.
- Fetzer, J.H.: 1990, *Artificial Intelligence: Its Scope and Limits*, Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Fodor, J.: 1975, *The Language of Thought*, Cambridge, MA: MIT Press.
- Fodor, J.: 1980, 'Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology'. Reprinted in Haugeland, J. (ed.), *Mind Design*, Cambridge, MA: MIT Press, pp. 307–338.
- Fodor, J.: 1987, *Psychosemantics*, Cambridge, MA: MIT Press.
- Haugeland, J.: 1981, 'Semantic Engines: An Introduction to *Mind Design*', in Haugeland, J. (ed.), *Mind Design*, Cambridge, MA: MIT Press, pp. 1–34.
- Haugeland, J.: 1985, *Artificial Intelligence: The Very Idea*, Cambridge, MA: MIT Press.
- Hempel, C.G.: 1949a, 'On the Nature of Mathematical Truth', in Feigl, H. and Sellars, W. (eds.) *Readings in Philosophical Analysis*, New York: Appleton-Century-Crofts, Inc., pp. 222–237.
- Hempel, C.G.: 1949b, 'Geometry and Empirical Science', in Feigl, H. and Sellars, W. (eds.), *Readings in Philosophical Analysis*, New York: Appleton-Century-Crofts, Inc., pp. 238–249.
- Hempel, C.G.: 1962, 'Rational Action', reprinted in Care, N. and Landesman, C. (eds.), *Readings in the Theory of Action*, Bloomington, IN: Indiana University Press, pp. 281–305.
- Maloney, C.: 1988, 'In Praise of Narrow Minds: The Frame Problem', in Fetzer, J. (ed.), *Aspects of Artificial Intelligence*, Dordrecht, The Netherlands: Kluwer Academic Publishers, pp. 55–80.
- Morris, C.: 1938, *Foundations of the Theory of Signs*, Chicago, IL: University of Chicago Press.
- Newell, A. and Simon, H.: 1976, 'Computer Science as Empirical Inquiry: Symbols and Search', reprinted in Haugeland, J. (ed.), *Mind Design*, Cambridge, MA: MIT Press, pp. 35–66.
- Popper, K.R.: 1978, 'Natural Selection and the Emergence of Mind', *Dialectica* 32, 339–355.
- Popper, K.R.: 1982, 'The Place of Mind in Nature', in Elvee, R. (ed.), *Mind in Nature*, San Francisco, CA: Harper & Row, Publishers, pp. 31–59.
- Putnam, H.: 1975, 'The Meaning of "Meaning"', in Gunderson, K. (ed.), *Language, Mind and Knowledge*, Minneapolis, MN: University of Minnesota Press, pp. 131–193.
- Quine, W.V.O.: 1960, *Word and Object*, Cambridge, MA: MIT Press.

- Quine, W.V.O.: 1975, 'Mind and Verbal Dispositions', in Guttenplan, S. (ed.), *Mind and Language*, Oxford, UK: Oxford University Press, pp. 83–95.
- Ryle, G.: 1949, *The Concept of Mind*, London, UK: Hutchinson.
- Skinner, B.F.: 1957, *Verbal Behavior*, New York: Appleton-Century-Crofts, Inc.
- Stich, S.: 1983, *From Folk Psychology to Cognitive Science*, Cambridge, MA: MIT Press.
- Winograd, T.: 1983, *Language as a Cognitive Process*, Vol. I, Reading, MA: Addison-Wesley.

University of Minnesota, Duluth

SELECTED BIBLIOGRAPHY

GENERAL BACKGROUND

- Anderson, A.: 1984, *Cognitive Psychology and Its Implications*, San Francisco, UK: Freeman.
- Armstrong, D.M.: 1968, *A Materialist Theory of Mind*, London, UK: Routledge and Kegan Paul.
- Armstrong, D.M.: 1977, 'The Causal Theory of the Mind', *Neue Heft für Philosophie* 11, Vandenhoek and Ruprecht, pp. 82–95.
- Bechtel, W.: 1987, 'Connectionism and the Philosophy of Mind: An Overview', *Southern Journal of Philosophyn*, Supp. 26, 17–41.
- Block, N. (ed.): 1980, *Readings in the Philosophy of Psychology*, Vol. I, Cambridge, MA: Harvard University Press.
- Block, N. (ed.) 1981, *Readings in the Philosophy of Psychology; Imagery*, Vol. II, Cambridge, MA: Harvard University Press.
- Bunge, M. and Ardila, R.: 1987, *Philosophy of Psychology*, New York, NY: Springer Verlag.
- Cole, D. and Foelber, R. (1984), 'Contingent Materialism', *Pacific Philosophical Quarterly* 65 (January), 74–85.
- Council for Philosophical Studies: 1983, *Psychology and the Philosophy of Mind in the Philosophy Curriculum*, San Francisco, CA: The Council for Philosophical Studies, San Francisco State University.
- Cummins, R.: 1975, 'Functional Analysis', *Journal of Philosophy* 72 (November 20, 1975), 741–765.
- Cummins, R.: 1983, *Psychological Explanation*, Cambridge, MA: MIT Press/Bradford Books.
- Davidson, D. and Harman, G. (eds.): 1971, *Semantics of Natural Language*, Dordrecht, The Netherlands: Reidel.
- Davidson, D.: 1975, 'Thought and Talk', in Guttenplan, S. (ed.): *Mind and Language*, Oxford, UK: Oxford University Press.
- Dennett, D.: 1969, *Content and Consciousness*, London, UK: Routledge & Kegan Paul.
- Dennett, D.: 1975, 'Why the Law of Effect won't Go Away', *Journal for the Theory of Social Behavior* 2, 169–187.
- Dennett, D.: 1978a, *Brainstorms: Philosophical Essays on Mind and Psychology*, Montgomery, VT: Bradford Books.
- Dennett, D.: 1978b, 'Skinner Skinned', in Dennett (1978a), pp. 53–70.
- Dennett, D.: 1984, *Elbow Room*, Cambridge, MA: MIT Press/Bradford Books.

- Dennett, D.: 1987, *The Intentional Stance*, Cambridge, MA: MIT Press.
- Descartes, R.: 1641, Latin edition, *Meditations on First Philosophy*; in Haldane, E.S. and Ross, G.R.T., eds. (1968).
- Devitt, M. and Sterelny, K.: 1987, *Language and Reality: An Introduction to the Philosophy of Language*, Cambridge, MA: Bradford Books, MIT Press.
- Dummett, M.A.: 1975, 'What is a Theory of Meaning?' in Guttenplan, S. (ed.): *Mind and Language*, Oxford, UK: Oxford University Press.
- Fetzer, J. (ed.): 1988, *Aspects of Artificial Intelligence*, Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Flanagan, O.J.: 1984, *The Science of the Mind*, Cambridge, MA: MIT Press/ Bradford.
- Fodor, J.: 1968, *Psychological Explanation: An Introduction to the Philosophy of Psychology*, New York, NY: Random House.
- Fodor, J.: 1974, 'Special Sciences, or The Disunity of Science as a Working Hypothesis', *Synthese* 28, 67–115.
- French, P.A. et al. (eds.): 1986, *Studies in the Philosophy of Mind: Midwest Studies in Philosophy*, Vol. X, Minneapolis, MN: University of Minnesota Press.
- Gregory, R.L. (ed.): 1987, *The Oxford Companion to the Mind*, London, UK: Oxford University Press.
- Gunderson, K. (ed.): 1975, *Language, Mind and Knowledge*, Minneapolis, MN: University of Minnesota Press.
- Guttenplan, S. (ed.): 1975, *Mind and Language*, Oxford, UK: Oxford University Press.
- Haldane, E., and Ross, G. (eds.): 1968, *The Philosophical Works of Descartes*, Vols. 1 and 2, Cambridge, UK: Cambridge University Press.
- Haugeland, J. (ed.): 1981, *Mind Design*, Cambridge, MA: MIT Press/Bradford Books.
- Hempel, C.C.: 1966, *Philosophy of Natural Science*, Englewood Cliffs, NJ: Prentice-Hall.
- Hofstadter, D. and Dennett, D.: 1981, *The Minds I*, New York, NY: Basic Books.
- James, W.: 1904, 'Does Consciousness Exist?' reprinted in James, W.: 1976, *Essays in Radical Empiricism*, Cambridge, MA: Harvard University Press.
- Johnson-Laird, P. and Wason, P. (eds.): 1977, *Thinking: Readings in Cognitive Science*, Cambridge, UK: Cambridge University Press.
- Leibniz, G.W. von: 1965, *Monadology*, Schrecker, P., trans., Indianapolis, IN: Bobbs-Merrill.
- Locke, J.: 1959, *Essay on Human Understanding*, New York, NY: Dover Publications.
- Lycan, W.: 1987, *Consciousness*, Cambridge, MA: MIT Press.
- Martindale, C.: 1981, *Cognition and Consciousness*, Homewood, IL: The Dorsey Press.
- McGinn, C.: 1982, *The Character of Mind*, Oxford, UK: Oxford University Press.
- Minsky, M.: 1987, *The Society of Mind*, New York, NY: Simon and Schuster.
- Nagel, T.: 1979, *Mortal Questions*, London, UK: Cambridge University Press.
- Neisser, U.: 1966, *Cognitive Psychology*, New York, NY: Appleton-Century-Crofts.
- Norman, D.A. (ed.): 1981, *Perspectives on Cognitive Science*, Hillsdale, NJ: Lawrence Erlbaum Associates.
- Nozick, R.: 1981, *Philosophical Explanations*, Cambridge, MA: Harvard University Press.

- Otto, H. and Tuedio, J. (eds.): 1988, *Perspectives on Mind*, Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Putnam, H.: 1987, *The Many Faces of Realism: The Paul Carus Lectures*, La Salle, PA: Open Court.
- Rorty, R.: 1979, *Philosophy and the Mirror of Nature*, Princeton, NJ: Princeton University Press.
- Rorty, R.: 1982, 'Contemporary Philosophy of Mind', *Synthese* 53, 323–48.
- Ryle, G.: 1949, *The Concept of Mind*, London, UK: Hutchinson.
- Savage, C.W. (ed.): 1978, *Perception and Cognition: Issues in the Foundations of Psychology* in *Minnesota Studies in the Philosophy of Science* Vol. IX, Minneapolis, MN: University of Minnesota Press.
- Sellars, W.: 1963, *Science, Perception and Reality*, London, UK: Routledge & Kegan Paul.
- Schiffer, S.: 1987, *Remnants of Meaning*, Cambridge, MA: MIT Press.
- Shaffer, J.: 1968, *Philosophy of Mind*, Englewood Cliffs, NJ: Prentice-Hall.
- Shoemaker, S.: 1981, 'Varieties of Functions', *Philosophical Topics*.
- Skinner, B.F.: 1933, *Science and Human Behavior*, New York, NY: MacMillan.
- Stalnaker, R.C.: 1984, *Inquiry*, Bradford Books, Cambridge, MA: MIT Press.
- Tienson, J.: 1987, 'Introduction to Connectionism', *Southern Journal of Philosophy*, Supp. 26, 1–16.
- Van Gulick, R.: 1988, 'A Functionalist Plea for Self-Consciousness', *Philosophical Review* 97, 149–181.
- White, S.L.: 1986, 'Curse of the Qualic', *Synthese* 68, 333–368.
- Williams, B.: 1978, *Descartes: The Project of Pure Inquiry*, New York, NY and London, UK: Penguin Books.
- Wittgenstein, L.: 1953, *Philosophical Investigations*, Oxford, UK: Basil Blackwell.
- Wittgenstein, L.: 1961, (orig. 1921), *Tractatus Logico-Philosophicus*, London, UK: Routledge & Kegan Paul.

I. COMPUTATIONAL CONCEPTIONS

- Anderson, A.: 1964, *Minds and Machines*, Englewood Cliffs, NJ: Prentice-Hall.
- Boden, M.: 1977, *Artificial Intelligence and Natural Man*, New York, NY: Basic Books.
- Boden, M.: 1981, *Minds and Mechanisms: Philosophical Psychology and Computational Models*, Ithaca, NY: Cornell University Press.
- Brand, M. and Harnish, R.M. (eds.): 1986, *The Representation of Knowledge and Belief*, Tucson, AZ: University of Arizona Press.
- Cole, D.: 1984, 'Thought and Thought Experiments', *Philosophical Studies* 45, 431–444.
- Creary, L.G. and Pollard, C.J.: 1985, 'A Computational Semantics for Natural Language', in *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, Chicago, IL: Association for Computational Linguistics, pp. 172–179.
- Dennett, D.: 1978a, 'Toward a Cognitive Theory of Consciousness' in *Minnesota Studies in the Philosophy of Science*, Vol. IX, Minneapolis, MN: University of

- Minnesota Press, pp. 201–228.
- Dennett, D.: 1978b, 'Why You Can't Make a Computer that Feels Pain', *Synthese* 38, 415–456.
- Dennett, D.: 1979, 'Artificial Intelligence as Philosophy and as Psychology', *Philosophical Perspectives on Artificial Intelligence*, New York, NY: Humanities Press.
- Dennett, D.: 1985, 'Cognitive Wheels: The Frame Problem of AI', in Hookway, C., ed., *Minds, Machines and Evolution: Philosophical Studies*, Cambridge, UK: Cambridge University Press.
- Dennett, D.: 1986, 'The Logical Geography of Computational Approaches (A View from the East Pole)', in Brand and Harnish (1986), pp. 59–79.
- Dennett, D.: 1988, 'When Philosophers Encounter Artificial Intelligence', *Daedalus* 117, 283–295.
- Dreyfus, H.L.: 1972, *What Computers Can't Do*, New York, NY: Harper and Row [second edition 1979].
- Dreyfus, H.L. and Haugeland, J.: 1974, 'The Computer as a Mistaken Model of the Mind', in S.C. Brown, ed., *Philosophy of Psychology*, London, UK: Macmillan, pp. 247–258.
- Fodor, J.: 1978, 'Tom Swift and His Procedural Grandmother', *Cognition* 6, 229–247.
- Fodor, J.: 1980a, 'Searle on What Only Brains Can Do', *The Behavioral and Brain Sciences* 3, pp. 431–432.
- Fodor, J.: 1980b, 'Methodological Solipsism Considered as a Research Strategy for Cognitive Psychology', *The Behavioral and Brain Sciences* 3, 63–73.
- Fodor, J.: 1983, *The Modularity of Mind*, Cambridge, MA: MIT Press/Bradford Books.
- Garfield, J. (ed.): 1987, *Modularity in Knowledge Representation and Natural Language Understanding*, Cambridge, MA: MIT Press.
- Gregory, R.: 1966, *Eye and Brain*, New York, NY: McGraw-Hill.
- Gregory, R.: 1970, *The Intelligent Eye*, London, UK.
- Gregory, R.: 1974, 'Perception as Hypothesis', in Brown, S.C. (ed.): *Philosophy of Psychology*, New York, NY: Harper and Row, pp. 195–210.
- Gunderson, K.: 1971, *Mentality and Machines*, New York, NY: Doubleday.
- Haugeland, J.: 1978, 'The Nature and Plausibility of Cognitivism', *The Behavioral and Brain Sciences* 2, 215–260.
- Haugeland, J.: 1981a, *Mind Design*, Cambridge, MA: MIT Press.
- Haugeland, J.: 1981b, 'Semantic Engines: An Introduction to Mind Design', in Haugeland, J. 1981a, pp. 1–34.
- Haugeland, J.: 1985, *Artificial Intelligence: The Very Idea*, Cambridge, MA: MIT Press.
- Hobbs, J. and Rosenschein, S.: 1987, 'Making Computational Sense of Montague's Intensional Logic', *Artificial Intelligence* 9, 287–306.
- Hubel, D.H., and Wiesel, T.N.: 1979, 'Brain Mechanisms of Vision', *Scientific American*, 241.
- Kosslyn, S.M., Pinker, S., Smith, G., and Schwartz, S.P.: 1979, 'On the Demystification of Mental Imagery', *The Behavioral and Brain Sciences* 2, 535–581.
- Johnson-Laird, P. and Wason, P. (eds.): 1977, *Thinking: Readings in Cognitive Science*, Cambridge, UK: Cambridge University Press.

- Laymon, R.: 1988, 'Some Computers Can Add (Even if the IBM 1620 Couldn't): Defending ENIAC's Accumulators Against Dretske', *Behaviorism* 16, 1-16.
- Lesperance, Y.: 1986, 'Towards a Computational Interpretation of Situation Semantics', *Computational Intelligence* 2, 9-27.
- Marr, D., and Poggio, T.: 1976, 'Cooperative Computation of Stereo Disparity', *Science* 194, 283-287.
- Marr, D.: 1982, *Vision: A Computational Investigation*. San Francisco, CA: W.H. Freeman & Co.
- Minsky, M.: 1973, 'Frame Systems: A Framework for Representing Knowledge', *The Society of Mind*, Cambridge, MA: MIT Artificial Intelligence Laboratory, pp. 243-272.
- Minsky, M.: 1975, 'Frame-System Theory', in Johnson-Laird and Wason (eds.), 1977.
- Minsky, M.: 1981a, 'A Framework for Representing Knowledge', in Haugeland, J. (ed.) 1981, pp. 95-128.
- Minsky, M.: 1981b, 'K-Lines: A Theory of Memory', in Norman, D. (ed.): *Perspectives on Cognitive Science*, Norwood, NJ: Ablex, pp. 87-103.
- Minsky, M.: 1982, 'Why People Think Computers Can't', *AI Magazine*, Fall.
- Moor, J.H.: 1988a, 'Testing Robots for Qualia', in Otto, H. and Tuedio, J. (eds.): 1988, *Perspectives on Mind*, Dordrecht, D. Reidel, pp. 107-118.
- Moor, J.H.: 1988b, 'The Pseudorealization Fallacy and the Chinese Room Argument', in Fetzer, J. (ed.): *Aspects of Artificial Intelligence*, Dordrecht, Netherlands: Kluwer Academic Publishers, pp. 35-53.
- Moore, R.C. and Hendrix, G.G.: 1982, 'Computational Models of Belief and the Semantics of Belief Sentences', in Peters, S. and Saarinen, E. (eds.): *Processes, Beliefs, and Questions*, Dordrecht, The Netherlands: D. Reidel, pp. 107-127.
- Newell, A. and Simon, H.: 1972, *Human Problem Solving*, Englewood Cliffs, NJ: Prentice-Hall.
- Newell, A. and Simon, H.: 1981, 'Computer Science as Empirical Inquiry: Symbols and Search', *Mind Design*, Montgomery, VT: Bradford, pp. 35-66.
- Newell, A.: 1982, 'The Knowledge Level', *Artificial Intelligence*, 18, 87-127.
- Norman, N. and Lindsay, P.: 1977, *Human Information Processing*, New York, NY: Harcourt Brace.
- O'Connor, J. (ed.): 1969, *Modern Materialism: Readings on Mind-Body Identity*, New York, NY: Harcourt Brace and World.
- Otto, H. and Tuedio, J. (eds.): 1988, *Perspectives on Mind*, Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Pollack, J.: 1988, 'My Brother, the Machine - Conception of Man as an Intelligent Machine', *Nous* 22, 173-212.
- Putnam, H.: The Mental Life of Some Machines', in O'Connor (ed.), 1969, pp. 263-281.
- Putnam, H.: 1960, 'Minds and Machines', in Hook, S. (ed.), *Dimensions of Mind*, New York, NY: New York University Press, pp. 138-164.
- Putnam, H.: 1964, 'Robots: Machines or Artificially Created Life?', *Journal of Philosophy* 61, 668-891.
- Putnam, H.: 1967, 'The Nature of Mental States', in Capitan, W.H., and Merrill, D.D. (eds.), *Art, Mind and Religion*, Pittsburgh, PA: University of Pittsburgh Press, pp. 150-161.

- Putnam, H.: 1985, 'Computational Psychology and Interpretation Theory' in B. Vermazen (ed.), *Essays on Davidson: Actions and Events*, Oxford, UK.
- Putnam, H.: 1988, 'Much Ado About Not Very Much', *Daedalus* 117 (Winter), 269–281.
- Pylyshyn, Z.: 1974, 'Minds, Machines, and Phenomenology: Some Reflections on Dreyfus' "What Computers Can't Do"', *Cognition* 3, 57–77.
- Pylyshyn, Z.: 1980, 'Computation and Cognition: Issues in the Foundations of Cognitive Science', *The Behavioral and Brain Sciences* 3, 111–132.
- Pylyshyn, Z.: 1981, 'The Imagery Debate: Analog Media Versus Tacit Knowledge', *Psychological Review*, 16–45.
- Pylyshyn, Z.: 1984, *Computation and Cognition: Toward a Foundation for Cognitive Science*, Cambridge, MA: Bradford/MIT Press.
- Pylyshyn, Z. (ed.): 1987, *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, Norwood, NJ: Ablex.
- Rapaport, W.J.: 1985, 'Machine Understanding and Data Abstraction in Searle's Chinese Room', *Proc. 7th Annual Conf. Cognitive Science Soc., University of California at Irvine 1985*, Hillsdale, NJ: Lawrence Erlbaum, pp. 341–345.
- Rapaport, W.J.: 1986a, 'Discussion: Searle's Experiments with Thoughts', *Philosophy of Science* 53, 271–279.
- Rapaport, W.J.: 1986b, 'Philosophy, Artificial Intelligence, and the Chinese-Room Argument', *Abacus* 3, 7–17.
- Rapaport, W.J.: 1988, 'Syntactic Semantics: Foundations of Computational Natural-Language Understanding', in Fetzer, J. (ed.), *Aspects of Artificial Intelligence*, Dordrecht, Netherlands: Kluwer Academic Publishers, pp. 81–131.
- Raphael, B.: 1976, *The Thinking Computer: Mind inside Matter*, San Francisco, CA: Freeman.
- Ringle, M. (ed.): 1979, *Philosophical Perspectives in Artificial Intelligence*, Atlantic Highlands, NJ: Humanities Press.
- Ringle, M.: 1982, 'Artificial Intelligence and Semantic Theory', in Simon, T.W. and Scholes R.J. (eds.), *Language, Mind, and Brain*, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 45–63.
- Schank, R.C. (ed.): 1975, *Conceptual Information Processing*, North-Holland, Amsterdam.
- Schank, R.C. and Riesbeck, C.K. (eds.): 1981, *Inside Computer Understanding: Five Programs Plus Miniatures*, Hillsdale, NJ: Lawrence Erlbaum Associates.
- Schank, R.: 1982, *Dynamic Memory*, Cambridge, UK: Cambridge University Press.
- Schank, R. and Abelson, R.: 1977, *Scripts, Plans, Goals, and Understanding*, Hillsdale, NJ: Erlbaum Associates.
- Searle, J.R.: 1980, 'Minds, Brains and Programs' in *The Behavioral and Brain Sciences* 3.
- Searle, J.R.: 1982, 'The Myth of the Computer', *New York Review of Books* (April 29, 1982), pp. 3–6.
- Searle, J.R.: 1987, 'Minds and Brains without Programs', in Blakemore, C. and Greenfield, S. (eds.), *Mindwaves: Thoughts on Intelligence, Identity, and Consciousness*, Oxford, UK: Basil Blackwell, pp. 209–233.
- Simon, H., 1974, *The Sciences of the Artificial*, Cambridge, MA: MIT Press.

- Sloman, A.: 1979, *The Computer Revolution in Philosophy*, Atlantic Highlands, NJ: Humanities Press.
- Stabler, E.P., Jr.: 1984, 'Berwick and Weinberg on Linguistics and Computational Psychology', *Cognition* 17, 155–179.
- Stich, S.C.: 1980, 'Paying the Price for Methodological Solipsism', *The Behavioral and Brain Sciences* 3, pp. 97–98.
- Stich, S.C.: 1983, *From Folk Psychology to Cognitive Science: The Case Against Belief*, Cambridge, MA: MIT Press/Bradford.
- Turing, A.M.: 1950, 'Computing Machinery and Intelligence', *Mind* 59, 433–460. Reprinted in Anderson, A.R. (ed.), *Minds and Machines*, Englewood Cliffs, NJ: Prentice-Hall, pp. 4–30.
- Ullman, S.: 1979, *The Interpretation of Visual Motion*, Cambridge, MA: MIT Press.
- Van Gulick, R.: 1988, 'Qualia, Functional Equivalence, and Computation' in Otto, H. and Tuorio, J. (eds.), *Perspectives on Mind*, Dordrecht, D. Reidel, pp. 119–126.
- Weizenbaum, J.: 1976, *Computer Power and Human Reason*, San Francisco, CA: Freeman.
- Wexler, K., and Culicover, P.: 1980, *Formal Principles of Language Acquisition*, Cambridge, MA: MIT Press.
- Winograd, T.: 1972, 'Understanding Natural Language', *Cognitive Psychology* 3, 1–191.
- Winograd, T.: 1981, 'What Does It Mean to Understand Language?', in Norman, D. (ed.), *Perspectives on Cognitive Science*, Norwood, NJ: Ablex, pp. 231–263.
- Winston, P.H., and Brown, R.H.: 1979, *Artificial Intelligence: An MIT Perspective*, Vols. I and II, Cambridge, MA: MIT Press.

II. CONNECTIONIST CONCEPTIONS

- Anderson, J.R.: 1985, *Cognitive Style* 9.
- Ballard, D.H.: 1986, 'Cortical connections and parallel processing: Structure and function', *Behavioral and Brain Sciences* 9, 150–175.
- Baron R.J.: 1986, *The Cerebral Computer: An Introduction to the Computational Structure of the Human Brain*, Hillsdale, NJ: Lawrence Erlbaum.
- Bechtel, W.: 1987, 'Connectionism and the Philosophy of Mind: An Overview', *Southern Journal of Philosophy* 26, Supplement, 17–41.
- Bechtel, W.: 1988, *Philosophy of Science: An Overview for Cognitive Science*, Hillsdale, NJ: Lawrence Erlbaum.
- Charniak, E.: 1987, 'Connectionism and Explanation', in *Proceedings of Theoretical Issues in Natural Language Processing* 3, Las Cruces, NM: New Mexico State University, pp. 68–72.
- Churchland, P.: 1981, 'Eliminative Materialism and Propositional Attitudes', *Journal of Philosophy* 78, 67–90.
- Churchland, P.S.: 1986, *Neurophilosophy*, Cambridge, MA: MIT Press.
- Clark, A.: 1987, 'Connectionism and Cognitive Science', in Hallam, J. and Mellish, C. (eds.), *Advances in Artificial Intelligence*, Chichester, UK: John Wiley.
- Cotrell, G.: 1987, 'Toward Connectionist Semantics', *Proceedings of Theoretical Issues in Natural Language Processing* 3, Las Cruces, NM: New Mexico State

- University, pp. 63–67.
- Cowan, J.D. and Sharp, D.: 1988, 'Neural Nets and Artificial Intelligence', *Daedalus* 117, 85–121.
- Dennett, D.: 1981, 'True Believers: The Intentional Strategy and Why it Works', *Scientific Explanation: Papers Based on Herbert Spencer Lectures Given in the University of Oxford*, Oxford, UK: The Clarendon Press.
- Dennett, D.: 1988, 'Fast Thinking', in Dennett, D. (ed.), *The Intentional Stance*, Cambridge, MA: Bradford Books/MIT Press.
- Derthick, M.A.: 1987, 'Counterfactual reasoning with direct models', *Proceedings of the Sixth National Conference on Artificial Intelligence*, Seattle, WA: University of Washington Press, pp. 346–351.
- Derthick, M.A.: 1987, 'A Connectionist Architecture for Representing and Reasoning about Structured Knowledge', *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*, Seattle, WA, pp. 131–142.
- Dreyfus, H. and Dreyfus, S.: 1988, 'Making a Mind versus Modelling the Brain: A.I. Back at a Branch Point', *Daedalus* 117, 1, (Winter), 15–43.
- Dreyfus, H.L. and Dreyfus, S.E.: 1986, *Mind Over Machine. The Power of Human Intuition and Expertise in the Era of the Computer*, New York, NY: Free Press.
- Farah, J.J.: 1988, 'Is Visual Imagery Really Visual? Overlooked Evidence from Neurophysiology', *Psychological Review* 95, 307–317.
- Feldman, J. and Ballard, D.: 1982, 'Connectionist Models and Their Properties', *Cognitive Science* 6, 205–254.
- Feldman, J.A.: 1981, 'A Connectionist Model of Visual Memory', *Parallel Models of Associative Memory*, Hinton, G.E. and Anderson, J.A. (eds.), Hillsdale, NJ: Lawrence Erlbaum.
- Field, H.: 1978, 'Mental Representation', in Block, N. (ed.), *Readings in the Philosophy of Psychology Vol. 2*, Cambridge, MA: Harvard University Press, pp. 78–114.
- Fodor, J.: 1987, 'Appendix: Why There Still Has To Be a Language of Thought', *Psychosemantics*, Cambridge, MA: MIT Press/Bradford Books, pp. 135–154.
- Fodor, J.: 1986, 'Information and Association', *Notre Dame Journal of Formal Logic* 27, 307–323.
- Fodor, J. and Pylyshyn, Z.: 1988, 'Connectionism and Cognitive Architecture: A Critical Analysis', *Cognition* 28, 3–71.
- Greeno, J.: 1987, 'The Cognition Connection', *New York Times Book Review* 28, January.
- Hillis, D.: 1985, *The Connection Machine*, Cambridge, MA: MIT Press.
- Hofstadter, D.R.: 1985, 'Waking up from the Boolean dream, or, subcognition as computation', *Metamagical Themas*, New York, NY: Basic Books.
- Hopfield, J.J.: 1982, 'Neural Networks and Physical Systems with Emergent Collective Computational Abilities', *Proc. Nat'l Academy of Sciences USA* 79(8), 2554–2558.
- Johnson, G.: 1987, 'The Latest Ideas about Brains, Minds and Bodies', *New York Times*, November 29, p. E18.
- Kosslyn, S.M.: 1980, *Image and Mind*, Cambridge, MA: Harvard University Press.
- Kosslyn, S.M. and Hatfield, G.: 1984, 'Representation Without Symbol Systems', *Social Research* 51, 1019–1054.

- Kripke, S.: 1980, *Wittgenstein on Following a Rule*, Cambridge, MA: Harvard University Press.
- Lycan, W.G.: 1987, *Consciousness*, Cambridge, MA: MIT Press/Bradford Books.
- Marr, D.: 1982, *Vision*, W.H. Freeman.
- McCauley, R.N.: 1988, 'Epistemology in an Age of Cognitive Science', *Philosophical Psychology* 1, 143-152.
- Minsky, M. and Papert, S.: 1969, *Perceptions*, Cambridge, MA: MIT Press.
- Newell, A.: 1980, 'Physical Symbol Systems', *Cognitive Science* 4, 135-183.
- Papert, S. (ed.): 1988a, *Daedalus* 117, (Winter). Special Issue on Artificial Intelligence.
- Papert, S.: 1988b, 'One AI or Many?' in Papert 1988a.
- Putnam, H.: 1960, 'Minds and Machines', Reprinted in Putnam 1975.
- Putnam, H.: 1966, 'The Mental Life of Some Machines', Reprinted in Putnam 1975.
- Putnam, H.: 1975, *Mind, Language and Reality: Philosophical Papers* (Vol. II), London, UK: Cambridge University Press.
- Putnam, H.: 1983, 'Computational Psychology and Interpretation Theory', *Philosophical Papers*, Vol. III: *Realism and Reason*, London, UK: Cambridge University Press, pp. 139-154.
- Pylyshyn, Z.W.: 1980, 'Cognition and Computation: Issues in the Foundations of Cognitive Science', *Behavioral and Brain Sciences* 3, 154-169.
- Pylyshyn, Z.W.: 1981, 'The imagery debate: Analogue Media Versus Tacit Knowledge', *Psychological Review* 88, 16-45.
- Pylyshyn, Z.W.: 1984a, *Computation and Cognition: Toward a Foundation for Cognitive Science*, Cambridge, MA: MIT Press Bradford Books.
- Pylyshyn, Z.W.: 1984b, 'Why Computation Requires Symbols', *Proceedings of the Sixth Annual Conference of the Cognitive Science Society*, Boulder, Colorado, August, 1984, Hillsdale, NJ: Lawrence Erlbaum.
- Reeke, G.N., Jr., and Edelman, G.: 1988, 'Real Brains and Artificial Intelligence', *Daedalus* 117 (1, Winter), 143-173.
- Rey, G.: 1983, 'Concepts and Stereotypes', *Cognition* 15, 237-262.
- Rey, G.: 1985, 'Concepts and Conceptions', *Cognition* 19, 297-303.
- Rumelhart, D.E., McClelland, J. and the PDP Research Group: 1986, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1, Cambridge, MA: MIT Press.
- Rumelhart, D.E., McClelland, J. and the PDP Research Group: 1986, *Foundations*, Vol. 2, Cambridge, MA: MIT Press.
- Rumelhart, D.E., McClelland, J. and the PDP Research Group: 1986, *Psychological and Biological Models*, Cambridge, MA: MIT Press.
- Rumelhart, D.E.: 1984, 'The Emergence of Cognitive Phenomena from Sub-Symbolic Processes', *Proceedings of the Sixth Annual Conference of the Cognitive Science Society*, Boulder, Colorado, August, 1984, Hillsdale, NJ: Erlbaum.
- Sampson, G.: 1987, 'Review of "Parallel Distributed Processing" by Rumelhart, McClelland, and PDP Group', *Language* 63 (4), 871-886.
- Schank, R.: 1975, *Conceptual Information Processing*, North-Holland.
- Schank, R.: 1982, *Dynamic Memory: A Theory of Learning in Computers and People*, Cambridge, UK: Cambridge University Press.

- Schank, R., and Abelson, R.: 1977, *Scripts, Plans, Goals and Understanding*, Hillsdale, NJ: John Wiley and Sons.
- Schneider, W.: 1987, 'Connectionism: Is It a Paradigm Shift for Psychology?', *Behavior Research Methods, Instruments, and Computers* 19, 73–83.
- Shastri, L. and Feldman, J.: 1985, 'The Constituent Structure of Connectionist Mental States: A Reply to Fodor and Pylyshyn', *Southern Journal of Philosophy*, Supplementary Volume.
- Skarda, C.A. and Freeman, W.: 1987, 'How Brains Make Chaos in Order to Make Sense of the World', *Behavioral and Brain Sciences* 10, 161–195.
- Smolensky, P.: 1986, 'Formal Modeling of Subsymbolic Processes: An Introduction to Harmony Theory', *Directions in the Science of Cognition*, Sharkey, N.E. (ed.), Ellis Horwood.
- Smolensky, P.: 1987a, 'Connectionist AI, Symbolic AI, and the Brain', *Artificial Intelligence Review* 1, 95–109.
- Smolensky, P.: 1987b, 'The Constituent Structure of Connectionist Mental States: A Reply to Fodor and Pylyshyn', *Southern Journal of Philosophy*, Supp. 26, 137–161.
- Smolensky, P.: 1988, 'On the Proper Treatment of Connectionism', *The Behavioral and Brain Sciences* 11, 1–74.
- Stabler, E.: 1985, 'How are Grammars Represented?', *Behavioral and Brain Sciences* 6, 391–420.
- Stich, S.: 1983, *From Folk Psychology to Cognitive Science*, Cambridge, MA: MIT Press.
- Thagard, P.: 1986, 'Parallel Computation and the Mind-Body Problem', *Cognitive Science* 10, 301–318.
- Touretzky, D. and Geva, S.: 1987, 'A Distributed Connectionist Representation for Concept Structures', *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*, pp. 155–164.

III. REPRESENTATIONAL CONCEPTIONS

- Anderson, J.R.: 1978, 'Arguments Concerning Representations for Mental Imagery', *Psychological Review* 85, 249–277.
- Block, N. (ed.): 1981, *Imagery*, Cambridge, MA: The MIT Press/Bradford Books.
- Brand, M. and Harnish, R.M. (eds.): 1986, *The Representation of Knowledge and Belief*, Tucson, AZ: University of Arizona Press.
- Bresnan, J. (ed.): 1982, *The Mental Representation of Grammatical Relations*, Cambridge, MA: MIT Press.
- Chomsky, N. and Fodor, J.: 1974, 'What the Linguist is Talking About', *Journal of Philosophy* 71 (12), 347–367. Reprinted in Block, N. (ed.), volume II.
- Chomsky, N. and Katz, J.J.: 1974, 'What the Linguist is Talking About', *Journal of Philosophy* 71, pp. 347–367. Reprinted in Block, N. (ed.), 1981, pp. 223–237.
- Chomsky, N.: 1980a, 'Rules and Representations', *The Behavioral and Brain Sciences* 3, 1–61.
- Chomsky, N.: 1980b, 'On Cognitive Structures and Their Development' in Piatelli-Palmerini, M. (ed.), *Language and Learning*, Cambridge, MA: Harvard University

- Press, pp. 35–82.
- Chomsky, N.: 1980c, *Rules and Representations*, New York, NY: Columbia University Press.
- Chomsky, N.: 1986, *Knowledge of Language: Its Nature, Origin, and Use*, New York, NY: Praeger Publishers.
- Cooper, D.: 1975, *Knowledge of Language*, New York, NY: Humanities Press.
- Dennett, D.: 1975, 'Brain-Writing and Mind Reading' in Gunderson, K. (ed.), *Language, Mind, and Knowledge: Minnesota Studies in the Philosophy of Science* VII, Minneapolis, MN: University of Minnesota Press. Reprinted in Dennett (1979), *Brainstorms*, pp. 39–52.
- Dennett, D.: 1977, 'Critical Notice: *The Language of Thought* by Jerry Fodor', *Mind* (April). Reprinted as 'A Cure for the Common Code?' in Dennett (1979), *Brainstorms*, pp. 90–108.
- Dennett, D.: 1982, 'Beyond Belief', in Woodfield, A. (ed.), *Thought and Object: Essays on Intentionality*, Oxford, UK: Oxford University Press, pp. 1–95.
- Devitt, M. and Sterelny, K.: 1987, *Language and Reality*, Cambridge, MA: MIT/Bradford Press.
- Dretske, F.: 1981, *Knowledge and the Flow of Information*, Cambridge, MA: Bradford Books/MIT Press.
- Dretske, F.: 1988, *Explaining Behavior: Reasons in a World of Causes*, Cambridge, MA: MIT Press.
- Evans, G.: 1981, 'Semantic Theory and Tacit Knowledge', in Holtzman, S. and Leich, C. (eds.), *Wittgenstein: To Follow a Rule*, London, UK: Routledge and Kegan Paul.
- Field, H.: 1977, 'Logic, Meaning and Conceptual Role', *Journal of Philosophy* 74, 379–409.
- Field, H.: 1978, 'Mental Representation', *Erkenntnis* 13, 9–61, reprinted in Block (1981).
- Fodor, J.: 1968, 'The Appeal to Tacit Knowledge in Psychological Explanation', *Journal of Philosophy* 65 (December 18), 627–640.
- Fodor, J.: 1975, *The Language of Thought*, New York, NY: Crowell.
- Fodor, J.: 1979, *The Language of Thought*, Cambridge, MA: Harvard University Press.
- Fodor, J.: 1981, *Representations: Philosophical Essays on the Foundations of Cognitive Science*, Cambridge, MA: MIT Press/Bradford Books.
- Fodor, J.: 1983, *The Modularity of Mind*, Cambridge, MA: MIT Press/Bradford Books.
- Fodor, J.: 1986, 'Off the Slippery Slope or Why Paramecia Don't Have Mental Representations', in French, P.A., et al. (eds.), *Studies in the Philosophy of Mind*, *Midwest Studies in Philosophy*, Vol. 10, Minneapolis, MN: University of Minnesota Press, pp. 3–23.
- Fodor, J.: 1987, *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*, Cambridge MA: Bradford Books, MIT Press.
- Fodor, J., Fodor, J.D., and Garrett, M.: 1975, 'The Psychological Unreality of Semantic Representations', *Linguistic Inquiry* 6, 515–531.
- Fodor, J. and Pylyshyn, Z.: 1981, 'How Direct is Visual Perception? Some Reflections on Gibson's "Ecological Approach"', *Cognition*, 9 (April), 139–196.

- Garfield, J.: 1988, *Belief in Psychology: A Study in the Ontology of Mind*, Cambridge, MA: MIT Press.
- Gibson, J.J.: 1966, *The Senses Considered as Perceptual Systems*, reprinted in Greenwood (1983).
- Gibson, J.J.: 1979, *The Ecological Approach to Visual Perception*, Hillsdale, NJ: L. Erlbaum Assoc.
- Goldman, A.I.: 1977, 'Perceptual Objects', *Synthese* 35, 257–284.
- Harman, G.: 1978, 'Is There Mental Representation', in Savage, C. (ed.), *Perception and Cognition: Issues in the Foundations of Psychology*, Minneapolis, MN: University of Minnesota Press, pp. 57–64.
- Harman, G.: 1982, 'Conceptual Role Semantics', *The Notre Dame Journal of Formal Logic*, 242–256.
- Haugeland, J.: 1978, 'The Nature and Plausibility of Cognitivism', *The Behavioral and Brain Sciences* 2, 215–225.
- Higgenbotham, J.: 1983, 'Is Grammar Psychological?', *Essays in Honor of Sidney Morgenbesser*, Indianapolis, IN: Hackett.
- Hills, D.: 1981, 'Introduction: Mental Representations and Languages of Thought' in Block, (ed.) (1981).
- Jackson, F.: 1977, *Perception*, London, UK: Cambridge University Press.
- Katz, J.J.: 1977, 'The Real Status of Semantic Representations', *Linguistic Inquiry* 8, 559–584. Reprinted in Block, N. (ed.), pp. 253–275.
- Kosslyn, S.M.: 1975, 'Information Representation in Visual Images', *Cognitive Psychology* 7.
- Kosslyn, S.M. *et al.*: 1978, 'On the Demystification of Mental Imagery', *The Behavioral and Brain Sciences* 2, 535–581.
- Kosslyn, S.M.: 1980a, *Image and Mind*, Cambridge, MA: Harvard University Press.
- Kosslyn, S.M.: 1980c, 'The Medium and the Message in Mental Imagery: a Theory', *Psychological Review* 88.
- Loar, B.: 1981, *Mind and Meaning*, Cambridge, UK: Cambridge University Press.
- Mehler, J. *et al.* (eds.): 1982, *Perspectives on Mental Representation*, Hillsdale, NJ: Erlbaum Press.
- Palmer, S.: 1981, 'Fundamental Aspects of Mental Representation' in Rosch, E.H., and Lloyd, B.B. (eds.), *Cognition and Categorization* Hillsdale, NJ: Erlbaum Press.
- Pyllyshyn, Z.: 1980b, 'The Imagery Debate: Analog Media vs. Tacit Knowledge', *Psychological Review* 88.
- Rock, I.: 1977, 'In Defense of Unconscious Inference' in Epstein, W. (ed.), *Stability and Constancy in Visual Perception*, New York, NY: John Wiley and Sons.
- Salmon, N.: 1986, *Frege's Puzzle*, Oxford, UK: Oxford University Press.
- Schwartz, R.: 1981, 'Imagery – There's More to it Than Meets the Eye', in Block, N. (ed.) (1981).
- Shepard, R.: 1978, 'The Mental Image', *American Psychologist* 33, 123–137.
- Shepard, R. and Cooper, L.: 1982, *Mental Images and Their Transformations*, Cambridge, MA: The MIT Press.
- Soames, S.: 1984, 'Semantics and Psychology', in Katz, J.J. (ed.), *The Philosophy of Linguistics*, Oxford, UK: Oxford University Press.

- Soames, S.: 1987, 'Semantics and Semantic Competence', in Schiffer, S. and Steele, S. (eds.), *Thought and Language: Second Arizona Colloquium on Cognitive Science*, Tucson, AZ: University of Arizona Press.
- Stalnaker, R.: 1976, 'Propositions', in Mackay, A. and Merrill, D. (eds.), *Philosophy of Language*, New Haven, CT: Yale University Press.
- Stampe, D.: 1977, 'Toward a Causal Theory of Linguistic Representation', in French, P., Uehling, T., and Wettstein, H. (eds.), *Midwest Studies in Philosophy Vol. II*, Minneapolis, MN: University of Minnesota Press, pp. 42–63.
- Stich, S.: 1971, 'What Every Speaker Knows', *Philosophical Review* 80, 476–496.
- Stich, S.: 1982, 'On the Ascription of Content', in Woodfield, A. (ed.), *Thought and Object: Essays on Intentionality*, Oxford, UK: Oxford University Press, pp. 153–206.
- Stich, S.: 1983, *From Folk Psychology to Cognitive Science: The Case Against Belief*, Cambridge MA: MIT Press.
- Tienson, J. and Horgan, T.: 1987, 'Settling into a New Paradigm', *Southern Journal of Philosophy*, Supp. 26, 97–113.
- Turvey, M.T.: 1977, 'Contrasting Orientations to the Theory of Visual Information Processing', *Psychological Review* 84, 67–88.
- Ullman, S.: 1980, 'Against Direct Perception', *The Behavioral and Brain Sciences* 3, 373–416.
- Van Gulick, R.: 1982, 'Mental Representation – A Functional View', *Pacific Philosophical Quarterly* 63, 3–20.

IV. MENTALITY AND INTENTIONALITY

- Brentano, F.: 1960, 'The Distinction between Mental and Physical Phenomena', *Realism and the Background of Phenomenology*, Glencoe, IL: Free Press, pp. 39–61.
- Chisholm, R.: 1967, 'On Some Psychological Concepts and the 'Logic' of Intentionality', in Castañeda, H.-N. (ed.), *Intentionality, Minds and Perception*, pp. 11–35.
- Cresswell, M.J.: 1985, *Structured Meanings: The Semantics of Propositional Attitudes*, Cambridge, MA: Bradford Books, MIT Press.
- Davidson, D.: 1970, 'Mental Events', in Foster, L. and Swanson, J.W. (eds.), *Experience and Theory*, Mass, pp. 79–101.
- Davidson, D.: 1974, 'Psychology as Philosophy', in Brown, S.C. (ed.), *Philosophy of Psychology*, London, UK: Macmillan, pp. 41–52.
- Dennett, D.: 1971, 'Intentional Systems', *Journal of Philosophy* 68, reprinted in Dennett (1978a).
- Dennett, D.: 1978, 'Toward A Cognitive Theory of Consciousness' in Savage (ed.), 1978.
- Dennett, D. and Haugeland, J.: 1987, 'Intentionality', in Gregory, R.L. (ed.), *Oxford Companion to the Mind*, Oxford, UK: Oxford University Press, pp. 383–386.
- Donnellan, K.S.: 1966, 'Reference and Definite Descriptions', *The Philosophical Review* 75, 281–304.

- Dretske, F.: 1980, 'The Intentionality of Mental States', in French, P. *et al.* (eds.) *Midwest Studies in Philosophy*, vol. V, pp. 281–294.
- Dreyfus, H. (ed.): 1982, *Husserl, Intentionality and Cognitive Science*, Cambridge, MA: MIT Press/Bradford Books.
- Grandy, R. and Warner, R.: 1986a, 'Paul Grice: A View of His Work', in Grandy and Warner (1986b).
- Grandy, R. and Warner, R. (eds.): 1986b, *Philosophical Grounds of Rationality: Intentions, Categories, Ends*, Oxford, UK: Oxford University Press.
- Grice, H.P.: 1957, 'Meaning', *Philosophical Review* **66**, 377–388.
- Grice, H.P.: 1968, 'Utterer's Meaning, Sentence Meaning, and Word-Meaning', *Foundations of Language*, 4, 225–252.
- Grice, H.P.: 1969, 'Utterer's Meaning and Intentions', *Philosophical Review* **78**, 147–177.
- Grice, H.P.: 1975, 'Logic and Conversation', in Davidson, D. and Harman, G. (eds.), *The Logic of Grammar*, Encino, CA: Dickenson.
- Hill, C.S.: 1988, 'Intentionality, Folk Psychology, and Reduction' in Otto, H. and Tuedio, J. (eds.), *Perspectives on Mind*, Dordrecht, D. Reidel, pp. 169–182.
- Hintikka, J.: 1975, *The Intentions of Intentionality and Other New Models for Modalities*, Dordrecht, The Netherlands: D. Reidel.
- Kenny, A.: 1963, *Action, Emotion, and Will*, London, UK: Routledge and Kegan Paul.
- Lewis, D.K.: 1969, *Convention: A Philosophical Study*, Cambridge, MA: Harvard University Press.
- Lycan, W.G.: 1969, 'On Intentionality and the Psychological', *American Philosophical Quarterly* **6**, 305–312.
- Marras, A. (ed.): 1972, *Intentionality, Mind and Language*, Urbana, IL: University of Illinois Press.
- Miller, G.A., Galanter, E., and Pribram, K.H.: 1960, *Plans and the Structure of Behavior*, New York, NY: Holt.
- Nagel, T.: 1974, 'What is it Like to Be a Bat?' *Philosophical Review* **83**, 435–450.
- Nagel, T.: 1986, *The View From Nowhere*, Oxford, UK: Oxford University Press.
- Nelson, R.: 1988, 'Mechanism and Intentionality: The New World Knot', in Otto, H. and Tuedio, J. (eds.), *Perspectives on Mind*, Dordrecht, D. Reidel, pp. 131–158.
- Otto, H. and Tuedio, J. (eds.): 1988, *Perspectives on Mind*, Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Peters, R.S.: 1958, *The Concept of Motivation*, London, UK: Routledge and Kegan Paul.
- Schiffer, S.: 1981, 'Truth and The Theory of Content', in Parrett, H. and Bouveresse, J. (eds.), *Meaning and Understanding*, Berlin, GM: Walter de Gruyter, pp. 204–222.
- Schiffer, S.: 1982, 'Intention-Based Semantics', *Notre Dame Journal of Formal Logic* **23**, 119–156.
- Schiffer, S.: 1986, 'Stalnaker's Problem of Intentionality', *Pacific Philosophical Quarterly* **67**, 87–97.
- Searle, J.: 1981, 'The Intentionality of Intention and Action', in Norman, D. (ed.) *Perspectives on Cognitive Science*, Norwood, NJ: Ablex, pp. 207–230.
- Searle, J.R.: 1983, *Intentionality: An Essay in the Philosophy of Mind*, Cambridge, UK: Cambridge University Press.

- Searle, J.R.: 1985, *Minds, Brains and Science*, Cambridge, MA: Harvard University Press.
- Sellars, W.: 1968, *Science and Metaphysics*, New York, NY: Routledge and Kegan Paul, Humanities Press.
- Sellars, W.: 1964, 'Notes on Intentionality', *Journal of Philosophy* 61, 655–664.
- Smith, D.W., and McIntyre, R.: *Husserl and Intentionality*, Boston, MA: Reidel.
- Speigelberg, H.: 1960, *The Phenomenological Movement*, Vol. I, New York, NY: Harper and Row.
- Speigelberg, H.: 1960, *The Phenomenological Movement*, Vol. II, New York, NY: Harper and Row.
- Stich, S.: 1981, 'Dennett on Intentional Systems', *Philosophical Topics* 12.
- Strawson, P.F.: 1968, 'Intention and Convention in Speech Acts', *Philosophical Review* 73, 439–460.
- Tuedio, J.A.: 1988, 'Intentional Transaction as a Primary Structure of Mind', in Otto, J. and Tuedio, H. (eds.), *Perspectives on Mind*, Dordrecht, D. Reidel, pp. 183–198.
- Van Gulick, R.: 1988, 'A Functionalist Plea for Self-Consciousness', *The Philosophical Review* 97, 149–181.
- Woodfield, A. (ed.): 1982, *Thought and Object: Essays on Intentionality*, Oxford, UK: Oxford University Press.

V. EPISTEMOLOGY AND COGNITION

- Alston, W.: 1980, 'Level-Confusions in Epistemology' in French *et al.* (eds.): *Midwest Studies in Philosophy* Vol. V: Studies in Epistemology, Minneapolis, MN: University of Minnesota Press, pp. 135–150.
- Anderson, C.A., Lepper, M.R., and Ross, L.: 1980, 'The Perseverance of Social Theories: The Role of Explanation in the Persistence of Discredited Information', *Journal of Personality and Social Psychology*.
- Anderson, J.R., and Bower, G.H.: 1973, *Human Associative Memory*, Washington, DC: V.H. Winston and Sons.
- Austin, J.: 1970, 'A Plea for Excuses', *Philosophical Papers*, 2nd Edition, pp. 175–204.
- Bartlett, F.C.: 1932, *Remembering: A Study in Experimental and Social Psychology*, Cambridge, UK: Cambridge University Press.
- Bransford, J. and McCarrell, N.: 1975, 'A Sketch of a Cognitive Approach to Comprehension', in Weimor, W.B. and Palermo, D.S. (eds.), *Cognition and the Symbolic Processes*, Hillsdale, NJ: Lawrence Erlbaum, pp. 189–230.
- Campbell, D.: 1974, 'Evolutionary Epistemology', in Schilpp, P.A. (ed.), *The Philosophy of Karl Popper*, Vol. 1, LaSalle, IL: Open Court Publishing Co, pp. 413–463.
- Carnap, R.: 1950, *Logical Foundations of Probability*, Chicago, IL: University of Chicago Press.
- Carroll, L.: 1895, 'What the Tortoise Said to Achilles', *Mind*, N.S. 4, 278–280.
- Cavell, S.: 1979, *The Claim of Reason*, Oxford, UK: The Clarendon Press.
- Cherniak, C.: 1981, 'Feasible Inferences', *Philosophy of Science* 48, 248–268.

- Chipman, S., Segal, J., and Glaser, R.: 1983, *Thinking and Learning Skills*, Vol. 2, Hillsdale, NJ: Erlbaum Associates.
- Chisholm, R.: 1977, *Theory of Knowledge*, 2nd Edition, Englewood Cliffs, NJ: Prentice-Hall.
- Churchland, P.M.: 1981, 'Eliminative Materialism and Propositional Attitudes', *The Journal of Philosophy*, **78**, 67–90.
- Churchland, P.S.: 1980, 'Language, Thought, and Information', *Nous* **14**, 147–170.
- Craik, F.I.M., and Lockhart, R.S.: 1972, 'Levels of Processing: A Framework for Memory Research', *Journal of Verbal Learning and Verbal Behavior* **1**, 671–684.
- Dawkins, R.: 1976, 'The Selfish Gene', Oxford, UK: Oxford University Press.
- Dennett, D.: 1969, *Content and Consciousness*, New York, NY: Humanities Press.
- Dennett, D.: 1978, 'Two Approaches to Mental Images', *Brainstorm*, Montgomery, VT: Bradford Books.
- Devitt, M.: 1984, *Realism and Truth*, Princeton, NJ: Princeton University Press.
- Doyle, J.: 1979, 'A Truth Maintenance System', *Artificial Intelligence* **12**, 231–272.
- Dretske, F.: 1969, *Seeing and Knowing*, Chicago IL: University of Chicago Press.
- Dretske, F.: 1981, *Knowledge and the Flow of Information*, Cambridge, MA: Bradford/MIT Press.
- Evans, J.: 1982, *The Psychology of Deductive Reasoning*. London, UK: Routledge and Kegan Paul.
- Fodor, J.: 1975, *The Language of Thought*, New York, NY: Thomas Y Crowell Company.
- Fodor, J.: 1978, 'Propositional Attitudes', *The Monist* **61**, 501–523.
- Fodor, J.: 1980, 'Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology', *The Behavioral and Brain Sciences* **3**, 63–73.
- Fodor, J. and Pylyshyn, Z.: 1981, 'How Direct is Visual Perception? Some Reflections on Gibson's Ecological Approach', *Cognition* **9**, 139–196.
- French, P., Uehling, T. Jr., and Wettstein, H. (eds.): 1980, *Midwest Studies in Philosophy Vol. V: Studies in Epistemology*, Minneapolis, MN: University of Minnesota Press.
- Goldman, A.: 1976, 'Discrimination and Perceptual Knowledge', *The Journal of Philosophy*, **73**, 771–791.
- Goldman, A.: 1978, 'Epistemic: The Regulative Theory of Cognition', *The Journal of Philosophy* **75**, 509–523.
- Goldman, A.: 1978a, 'Epistemology and the Psychology of Cognition', *The Monist* **61**, 525–535.
- Goldman, A.: 1978b, 'Epistemic: The Regulative Theory of Cognition', *Journal of Philosophy*, **75**, 509–523.
- Goldman, A.: 1979, 'What is Justified Belief', in Pappas, G.S. (ed.), *Justification and Knowledge* Dordrecht, The Netherlands: D. Reidel, pp. 1–23.
- Goldman, A.: 1980, 'The Internalist Conception of Justification', in French, P., Uehling, T. and Wettstein, H. (eds.), *Midwest Studies in Philosophy, Volume V: Studies in Epistemology*, Minneapolis, MN: University of Minnesota Press, pp. 27–51.
- Goldman, A.: 1986, *Epistemology and Cognition*, Cambridge, MA: Harvard University Press.

- Goldstein, I. and Papert, S.: 1977, 'Artificial Intelligence, Language and the Study of Knowledge', *Cognitive Science* 1, 84–123.
- Gruber, H.: 1982, 'Genes for General Intellect Rather than for Particular Culture', *The Behavioral and Brain Sciences* 5, 11–12.
- Hamlyn, D.: 1967, 'Epistemology, History of', in Edwards, P. (ed.), *the Encyclopedia of Philosophy*, Vol. 3, New York, NY: Macmillan, pp. 8–38.
- Harman, G.: 1973, *Thought*, Princeton, NJ: Princeton University Press.
- Hintikka, J.: 1962, *Knowledge and Belief*, Ithaca, NY: Cornell University Press.
- Hirst, R.: 1959, *The Problems of Perception*, New York, NY: Humanities Press.
- Hollis, M. and Lukes, S. (eds.): 1982, *Rationality and Relativism*, Cambridge MA: MIT Press.
- Jeffrey, R.C.: 1956, *The Logic of Decision*, New York, NY: McGraw-Hill Inc.
- Kaplan, B.: 1971, 'Genetic Psychology, Genetic Epistemology, and Theory of Knowledge', in Theodore Mischel (ed.), *Cognitive Development and Epistemology*, New York, NY: Academic Press, pp. 61–81.
- Klatzky, R.: 1980, *Human Memory*, 2nd Edition, San Francisco, CA: W.H. Freeman and Company.
- Kuhn, T.: 1962, *The Structure of Scientific Revolutions*, Chicago, IL: University of Chicago Press.
- Loar, B.: 1987, 'Truth Beyond All Verification', in Taylor, B. (ed.), *Michael Dummett*, Amsterdam: Martinus Nijhoff, pp. 81–116.
- Loftus, E.: 1979, *Eyewitness Testimony*, Cambridge, MA: Harvard University Press.
- Lycan, W.: 1988, *Judgment and Justification*, Cambridge, UK: Cambridge University Press.
- Miller, G.: 1959, 'The Magical Number Seven, Plus or Minus Two: Some Limits in our Capacity for Processing Information', *Psychological Review* 63, 81–97.
- Miller, G. and Johnson-Laird, P.N.: 1976, *Language and Perception*, Cambridge, MA: Harvard University Press.
- Mischel, T. (ed.): 1971, *Cognitive Development and Epistemology*, New York, NY: Academic Press.
- Moravcsik, J.M.E.: 1981, 'How Do Words Get Their Meaning', *The Journal of Philosophy* 88, 5–24.
- Nisbett, R. and Ross, L.: 1980, *Human Inference: Strategies and Shortcomings of Social Judgement*, Englewood Cliffs, NJ: Prentice Hall.
- Nozick, R.: 1981, *Philosophical Explanations*, Cambridge, MA: Belknap/Harvard University Press.
- Pappas, G. (ed.): 1980, *Justification and Knowledge* Dordrecht, The Netherlands: D. Reidel.
- Pappas, G. and Swain, M. (eds.): 1978, *Essays on Knowledge and Justification* Ithaca, NY: Cornell University Press.
- Perkins, M.: 1983, *Sensing the World*, Indianapolis, IN: Hackett.
- Piaget, J.: 1970, *Genetic Epistemology*, New York, NY: Columbia University Press.
- Piaget, J.: 1971, *Biology and Knowledge: An Essay on the Relations Between Organic Regulations and Cognitive Processes*, Chicago, IL: University of Chicago Press.
- Pollock, J.: 1983, 'Epistemology and Probability', *Synthese* 55, 231–252.
- Posner, M.: 1973, *Cognition: An Introduction*, Glenview, IL: Scott Foresman.

- Powers, L.: 1978, 'Knowledge by Deduction', *The Philosophical Review* 87, 337-371.
- Putnam, H.: 1975a, 'The Meaning of "Meaning"', in Putnam, H., *Mind, Language and Reality, Philosophical Papers*, Vol. 2, New York, NY: Cambridge University Press, pp. 215-271.
- Putnam, H.: 1975b, *Mind, Language and Reality*, Cambridge, NY: Cambridge University Press.
- Putnam, H.: 1978, *Meaning and the Moral Sciences*, London, UK: Routledge and Kegan Paul.
- Putnam, H.: 1981, *Reason, Truth and History*, Cambridge, UK: Cambridge University Press.
- Putnam, H.: 1988, *Representation and Reality*, Cambridge, MA: MIT Press.
- Quine, W.V.O.: 1960, *Word and Object*, Cambridge, MA: The MIT Press.
- Quine, W.V.O.: 1969a, 'Epistemology Naturalized', in *Ontological Relativity and Other Essays*, New York, NY: Columbia University Press, pp. 69-90.
- Quine, W.V.O.: 1969b, 'Natural Kinds', in *Ontological Relativity and Other Essays*, New York, NY: Columbia University Press, pp. 114-138.
- Quine, W.V.O.: 1969c, 'Ontological Relativity', *Ontological Relativity and Other Essays*, New York, NY: Columbia University Press, pp. 26-68.
- Quine, W.V.O.: 1975, 'On Empirically Equivalent Systems of The World', *Erkenntnis* 9, 313-328.
- Rorty, R.: 1979, *Philosophy and the Mirror of Nature*, Princeton, NJ: Princeton University Press.
- Rorty, R.: 1979, 'Pragmatism, Relativism and Irrationalism', *Proceedings and Addresses of the American Philosophical Association* 53.
- Schmidt, F.: 1981, 'Justification as Reliable Indication or Reliable Process', *Philosophical Studies* 40, 409-417.
- Schmidtt, F.: 1984, 'Reliability, Objectivity, and the Background of Justification', *Australasian Journal of Philosophy* 62, 1-15.
- Shope, R.: 1983, *The Analysis of Knowing*, Princeton, NJ: Princeton University Press.
- Smith, E. and Medlin, D.: 1981, *Categories and Concepts*, Cambridge, MA: Harvard U.P.
- Sosa, E.: 1980, 'The Foundations of Foundationalism', *Nous* 14, 547-654.
- Swain, M.: 1981, *Reasons and Knowledge*, Ithaca, NY: Cornell University Press.
- Tolman, E.C.: 1948, 'Cognitive Maps in Rats and Men', *Psychological Review* 55, 189-208.
- Ullman, S.: 1980, 'Against Direct Perception', *The Behavioral and Brain Sciences* 3, 373-415.
- Wason, P.C. and Johnson-Laird, P.: 1972, *The Psychology of Reasoning*, London, UK: Batsford.

VI. THE MENTAL AND THE PHYSICAL

- Block, N.: 1978, 'Troubles with Functionalism', *Minnesota Studies in the Philosophy of Science*, Vol. IX.; reprinted in Block (ed.) (1980) *Readings in the Philosophy of Psychology*, Volume I, Cambridge, MA: Harvard University Press, pp. 268-305.

- Block, N.: 1979, 'Reductionism', in Reich, W. (ed.) *Encyclopedia of Bioethics*, New York, NY: Macmillan.
- Block, N. and Fodor, J.: 1972, 'What Psychological States are Not', *The Philosophical Review* 81, 159–181.
- Borst, C.V. (ed.): 1970, *The Mind-Brain Identity Theory*, New York, NY: Macmillan.
- Boyd, R.: 1980, 'Materialism without Reductionism: What Physicalism Does Not Entail', in Block, N. (ed.), *Readings in the Philosophy of Psychology*, Volume I, Cambridge, MA: Harvard University Press, pp. 67–106.
- Causey, R.: 1977, *The Unity of Science*, Dordrecht, The Netherlands: Reidel.
- Churchland, P.M.: 1970, 'The Logical Character of Action Explanations', *Philosophical Review* 79, 214–236.
- Churchland, P.M.: 1979, *Scientific Realism and the Plasticity of Mind*, Cambridge, UK: Cambridge University Press.
- Churchland, P.M.: 1981, 'Eliminative Materialism and Propositional Attitudes', *Journal of Philosophy* 78, 67–90.
- Churchland, P.M.: 1984, *Matter and Consciousness*, Cambridge, MA: MIT Press/Bradford Books.
- Churchland, P.M. and Churchland, P.S.: 1981, 'Functionalism, Qualia and Intentionality', *Philosophical Topics* 12.
- Churchland, P.M. and Churchland, P.S.: 1982, *Mind, Brain, and Function*, Norman, OK: University of Oklahoma Press.
- Churchland, P.S.: 1980, 'A Perspective on Mind-Brain Research', *Journal of Philosophy* 77, 185–207.
- Churchland, P.S.: 1983, 'Consciousness: The Transmutation of a Concept', *Pacific Philosophical Quarterly* 64, 80–95.
- Davidson, D.: 1967, 'Causal Relations', *Journal of Philosophy* 64, 691–703.
- Feldman, F.: 1980, 'Identity, Necessity and Events', in Block, N. (ed.) (1980), pp. 148–155.
- Feigl, H.: 1967, *The 'Mental' and the 'Physical': The Essay and a Postscript*, Minneapolis, MN: University of Minnesota Press.
- Fetzer, J.H. (ed.): 1988, *Aspects of Artificial Intelligence*, Dordrecht Netherlands: Kluwer Academic Publishers.
- Feyerabend, P.: 1963, 'Comment: "Mental Events and the Brain"', *Journal of Philosophy* 50, 295–269.
- Feyerabend, P.: 1963, 'Materialism and the Mind-Body Problem', *Review of Metaphysics* 17, 49–66.
- Field, H.: 1977, 'Logica, Meaning, and Conceptual Role', *Journal of Philosophy* 74, 379–409.
- Fodor, J.: 1981, 'The Mind-Body Problem', *Scientific American* 244, 124–133.
- Gazzaniga, M.S.: 1970, *The Bisected Brain*, New York, NY: Appleton.
- Globus, G. et al.: 1976, *Consciousness and the Brain*, New York, NY: Plenum.
- Glover, J. (ed.): 1976, *The Philosophy of Mind*, Oxford, UK: Oxford University Press.
- Guttenplan, S. (ed.): 1975, *Mind and Language*, Oxford, UK: Oxford University Press.
- Hempel, C.G.: 1970, 'Reduction: Ontological and Linguistic Facets', in Morgenbesser, S., Suppes, P. and White, M. (eds.), *Essays in Honor of Ernest Nagel*, New York, NY: St. Martin's Press, pp. 179–199.

- Kandel, E.R.: 1976, *The Cellular Basis of Behavior*, San Francisco, CA: Freeman.
- Kim, J.: 1966, 'On the Psycho-Physical Identity Theory', *American Philosophical Quarterly* 3, 227–235.
- Kim, J.: 1969, 'Events and Their Descriptions: Some Considerations', in Rescher, N. et al. (eds.), *Essays in Honor of C.G. Hempel*, Dordrecht, The Netherlands: Reidel, pp. 198–215.
- Kim, J.: 1982, 'Psychological Supervenience', *Philosophical Studies* 41, 51–70.
- Kripke, S.: 1971, 'Naming and Necessity' in Davidson, D. and Harman, G. (eds.), *Semantics and Natural Language*, Dordrecht, The Netherlands: Reidel, pp. 253–355.
- Lewis, D.: 1966, 'An Argument for the Identity Theory', *Journal of Philosophy* 63, 17–25, reprinted in Rosenthal, (ed.) (1971).
- Lewis, D.: 1972, 'Psychophysical and Theoretical Identifications', *Australasian Journal of Philosophy*, 50, 249–258.
- Lewis, D.: 1980, 'Mad Pain and Martian Pain', in Block, N. (1980a), pp. 216–222.
- Lewis, D.: 1983, *Philosophical Papers*, Vol. 1, Oxford, UK: Oxford University Press.
- Lycan, W.G.: 1973, 'Inverted Spectrum', *Ratio* 15, 315–319.
- Lycan, W.G.: 1974, 'Kripke and the Materialists', *Journal of Philosophy* 71, 677–689.
- Lycan, W.G.: 1980, 'Form, Function and Feel', *Journal of Philosophy* 78, 24–50.
- Marks, C.E.: 1980, *Commissurotomy, Consciousness and Unity of Mind*, Cambridge, MA: MIT Press and Bradford Books.
- Nagel, T.: 1971, 'Brain Bisection and the Unity of Consciousness', *Synthese* 22, 396–413, reprinted in Glover, J. (ed.) (1976).
- Nisbett, R. and Wilson, T.: 1977, 'Telling More Than We Can Know: Verbal Reports on Mental Processes', *Psychological Review* 84.
- O'Connor, J. (ed.): 1969, *Modern Materialism: Readings on Mind-Body Identity*, New York, NY: Harcourt Brace and World.
- Owens, J.: 1983, 'Functional and Propositional Attitudes', *Nous* 17, 529–549.
- Parret, H. and Bouveresse, J. (eds.): 1981, *Meaning and Understanding*, Berlin, GM: New York, NY: de Gruyter.
- Place, U.T.: 1956, 'Is Consciousness a Brain Process?', *British Journal of Psychology* 47, reprinted in O'Connor (ed.), 1969, pp. 21–31.
- Place, U.T.: 1988, 'Thirty Years On – Is Consciousness Still a Brain Process?', *Australasian Journal of Philosophy* 66, 208–219.
- Popper, K. with Eccles, J.C.: 1977, *The Self and Its Brain*, New York, NY: Springer-Verlag.
- Puccetti, R.: 1973, 'Brain Bisection and Personal Identity', *British Journal of Philosophy of Science* 24, 330–355.
- Puccetti, R. and Dykes, R.: 1978, 'Sensory Cortex and the Mind-Brain Problem', *The Behavioral and Brain Sciences* 1, pp. 337–344.
- Quine, W.V.O.: 1975, 'Mind and Verbal Dispositions', in Guttenplan, S. (ed.) (1975), pp. 83–95.
- Rey, G.: 1980, 'Functionalism and the Emotions' in Rorty, R. (ed.) (1980).
- Rorty, R.: 1965, 'Mind-Body Identity, Privacy and Categories', *The Review of Metaphysics* 19, 24–54.
- Rorty, R.: 1970, 'In Defense of Eliminative Materialism', *Review of Metaphysics* 24, 112–121.

- Rosenthal, D.M. (ed.): 1971, *Materialism and the Mind-Body Problem*, Englewood Cliffs, NJ: Prentice-Hall.
- Shaffer, J.A.: 1977, 'Personal Identity: The Implications of Brain Bisection and Brain Transplants', *Journal of Medicine and Philosophy* 2, 147-161.
- Shoemaker, S.: 1975, 'Functionalism and Qualia', *Philosophical Studies* 27, 291-315.
- Shoemaker, S.: 1982, 'The Inverted Spectrum', *Journal of Philosophy* 79, 357-381.
- Smart, J.J.C.: 1959, 'Sensations and Brain Processes', *Philosophical Review* 68, 141-156.
- Sperry, R.W.: 1977, 'Forebrain Commissurotomy and Conscious Awareness', *Journal of Medicine and Philosophy* 2, 101-126.
- Stalnaker, R.: 1984, *Inquiry*, Cambridge MA: MIT Press, a Bradford Book.
- Taylor, C.: 1967, 'Mind-Body Identity: A Side Issue?', *Philosophical Review* 76, 206-207.
- Van Gulick, R.: 1988, 'A Functionalist Plea for Self-Consciousness', *The Philosophical Review* 97, 149-181.
- Vendler, Z.: 1972, *Res Cogitans*, New York, NY: Cornell University Press.
- Von Eckardt, B.: 1978, 'Inferring Functional Localization from Neurological Evidence' in Walker, E. (ed.) *Explorations in the Biology of Language*, Vt: Bradford, pp. 27-66.
- White, S.L.: 1986, 'Curse of the Qualia', *Synthese* 68, 333-368.
- Wilkes, K.V.: 1978, 'Physicalism', London, UK: Routledge and Kegan Paul.
- Wilkes, K.V.: 1981, 'Functionalism, Psychology and the Philosophy of Mind', *Philosophical Topics* 12.

VII. DISPOSITIONAL CONCEPTIONS

- Alston, W.: 1971, 'Dispositions and Occurrences', *Canadian Journal of Philosophy* 1, 125-154.
- Armstrong, D.M.: 1969, 'Dispositions Are Causes', *Analysis* 30, 23-26.
- Armstrong, D.M.: 1973, *Belief, Truth, and Knowledge*, New York, NY: Cambridge University Press.
- Ayer, A.J.: 1936, *Language, Truth, and Logic*, reprint NY: Dover 1952.
- Ayer, A.J.: 1963, *The Concept of a Person and other Essays*, London, UK.
- Block, N. and Fodor, J.: 1972, 'What Psychological States are Not', *Philosophical Review* 31, 159-181.
- Carnap, R.: 1936, 'Testability and Meaning (I)', *Philosophy of Science* 3, 419-471.
- Carnap, R.: 1937, 'Testability and Meaning (II)', *Philosophy of Science* 4, 1-40.
- Carnap, R.: 1956, 'The Methodological Character of Theoretical Concepts', in Fiegl, H. and Scriven, M. (eds.), *Minnesota Studies in the Philosophy of Science*, Vol. 1, Minneapolis, MN: University of Minnesota Press, pp. 38-76.
- Carnap, R.: 1963, 'Replies and Systematic Expositions', in Schilpp, P. (ed.), *The Philosophy of Rudolf Carnap*, La Salle, IL: Open Court Publishing Company, pp. 889-1013.
- Chomsky, N.: 1959, 'A Review of B.F. Skinner's *Verbal Behavior*', *Language* 35, 26-58.

- Dennett, D.: 1969, *Content and Consciousness*, London, UK: Routledge & Kegan Paul.
- Dennett, D.: 1978, 'Skinner Skinned', in Dennett, D. (ed.), *Brainstorms*, Montgomery, VT: Bradford.
- Fetzer, J.H.: 1977, 'A World of Dispositions', *Synthese* 34, 397–421.
- Fetzer, J.H.: 1978, 'On Mellor on Dispositions', *Philosophia* 7, 651–660.
- Fetzer, J.H.: 1981, *Scientific Knowledge*, Dordrecht, The Netherlands: D. Reidel.
- Fetzer, J.H.: 1984, 'Reduction Sentence "Meaning Postulates"', in Rescher, N. (ed.), *The Heritage of Logical Positivism*, Lanham, MD: University Press of America, pp. 55–65.
- Fetzer, J.H.: 1988a, 'Mentality and Creativity', *Journal of Social and Biological Structures* 11, 82–85.
- Fetzer, J.H.: 1988b, 'Signs and Minds: An Introduction to the Theory of Semiotic Systems', in Fetzer, J. (ed.), *Aspects of Artificial Intelligence*, Dordrecht, The Netherlands: Kluwer Academic Publishers, pp. 133–161.
- Fetzer, J.H.: 1990, *Artificial Intelligence: Its Scope and Limits*, Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Hempel, C.G.: 1952, *Fundamentals of Concept Formation in Empirical Science*, Chicago, IL: University of Chicago Press.
- Hempel, C.G.: 1965, *Aspects of Scientific Explanation*, New York, NY: The Free Press.
- Hempel, C.: 1949, 'The Logical Analysis of Psychology' in Feigl, H. and Sellars, W. (eds.), *Readings in Philosophical Analysis*, New York, NY: Appleton-Century-Crofts, pp. 373–384.
- Hull, C.L.: 1943, *Principles of Behavior*, New York, NY: Appleton-Century-Crofts.
- Kripke, S.: 1982, *Wittgenstein on Rules and Representations*, Cambridge, MA: Harvard University Press.
- Levi, I. and Morgenbesser, S.: 1964, 'Belief and Disposition', *American Philosophical Quarterly*, 1, 221–231.
- Mellor, D.H.: 1974, 'In Defense of Dispositions', *Philosophical Review* 83, 157–181.
- Pap, A.: 1959, 'Dispositional Concepts and Extensional Logic', in Feigl, H., Scriven, M. and Maxwell, G. (eds.), *Minnesota Studies in the Philosophy of Science*, Vol. II, Minneapolis, MN: University of Minnesota Press, pp. 196–224.
- Pap, A.: 1963, 'Reduction Sentences and Dispositional Concepts', in Schilpp, P. (ed.), *The Philosophy of Rudolf Carnap*, La Salle, IL: Open Court Publishing Company, pp. 589–597.
- Prior, E.: 1985, *Dispositions*, Aberdeen, Scotland: University of Aberdeen Press.
- Putnam, H.: 1965, 'Brains and Behavior', in Butler, R.J. (ed.), *Analytical Philosophy*, Vol. 2, Oxford, UK: Blackwell, 1965, pp. 1–19.
- Quine, W.V.O.: 1960, *Word and Object*, Cambridge, MA: Harvard University Press.
- Quine, W.V.O.: 1975, 'Mind and Verbal Dispositions', in Guttenplan, S. (ed.), *Mind and Language*, Oxford, UK: Oxford University Press, pp. 83–95.
- Rozeboom, W.W.: 1973, 'Dispositions Revisited', *Philosophy of Science* 40, 59–74.
- Ryle, G.: 1949, *The Concept of Mind*, London, UK: Hutchinson.
- Skinner, B.F.: 1938, *The Behavior of Organisms*, New York, NY: Appleton-Century Crofts.

- Skinner, B.F.: 1953, *Science and Human Behavior*, New York, NY: Macmillan.
- Skinner, B.F.: 1957, *Verbal Behavior*, New York, NY: Appleton-Century-Crofts, Inc.
- Skinner, B.F.: 1972, *Contingencies of Reinforcement*, New York, NY: Appleton-Century-Crofts, Inc.
- Skinner, B.F.: 1974, *About Behaviorism*, New York, NY: Random House.
- Stalnaker, R.: 1984, *Inquiry*, Cambridge, MA: MIT Press.
- Stevenson, C.L.: 1940, *Ethics and Language*, New Haven, CT: Yale University Press.
- Tolman, E.C.: 1932, *Purposive Behavior in Animals and Men*, New York, NY: Appleton-Century Crofts.
- Tuomela, R.: 1977, 'Dispositions, Realism, and Explanation', *Synthese* 34, 457-478.
- Tuomela, R. (ed.): 1978, *Dispositions*, Dordrecht, The Netherlands: D. Reidel.
- Wittgenstein, L.: 1953, *Philosophical Investigations*, Oxford, UK: Basil Blackwell.

INDEX OF NAMES

- Abelson, R. 408, 411
Ackley, D. 187, 191, 202
Adams, F. 87
Akins, K. 200
Albritton, R. 67
Alston, W. 67, 68, 417, 423
Anderson, A. 68, 142, 143, 366, 403, 405
Anderson, C. 342, 417
Anderson, J. 202, 146, 155, 156, 157, 159, 187, 189, 203, 342, 409, 412, 417
Anscombe, G. 50, 68
Arbib, M. 68, 196
Ardilla, R. 403
Armstrong, D. 11–12, 54, 59, 68, 359, 361, 362, 363, 373, 403, 423
Arnheim, R. 18, 91, 113
Asanuma, C. 166, 202
Asquith, P. 143, 144
Augustine 31
Austin, J. 50, 66, 68, 288, 303, 417
Ayala, F. 373
Ayers, A. 423
Ballard, D. 146, 184, 200, 201, 202, 203, 409, 410
Baron, R. 409
Bartlett, F. 341, 342, 417
Bechtel, W. xi, 12, 41–43, 373, 403, 409
Berkowitz, L. 344
Bernard, C. 370, 373
Black, M. 69
Block, N. 40, 67, 68, 80, 88, 113, 141, 142, 250, 280, 349, 358, 361, 372, 373, 403, 412, 420, 423
Bobrow, D. 204
Boden, M. 68, 405
Boltzmann, H. 187, 191, 192, 201, 202, 203
Borst, C. 55, 68, 280, 420
Bouveresse, J. 422
Bower, G. 142, 143, 336, 342
Boyd, R. 420
Braithwaite, R. 280
Brand, M. 405, 412
Bransford, J. 337, 342, 417
Brentano, F. 4
Bresnan, J. 412
Bretano, F. 63–64, 415
Brewer, W. 107, 113, 204
Broadbent, D. 142, 143
Brooks, B. 88n
Brown, C. 203
Brown, R. 409
Bruce, B. 204
Buchler, J. 392, 394, 401
Bunge, M. 403
Burge, T. 88, 209
Butler, R. 70
Campbell, D. 34, 307, 342, 370, 373, 417
Carnap, R. 290, 298, 340, 342, 379, 401, 417, 423
Carroll, L. 342, 417
Castañeda, H. 68, 71, 266
Causey, R. 420
Cavell, S. 292, 294, 303, 417
Changeux, J. 205
Charniak, E. 378, 401, 409
Chase, W. 144, 342
Cherniak, C. 143, 323, 342, 412, 417
Chihara, C. 54, 68

- Chipman, S. 417
 Chisholm, R. xi, 13, 29–31, 32, 64,
 68, 266, 310, 341, 415, 417
 Chomsky, N. 61, 107, 412, 413, 423
 Church, A. 150, 161
 Churchland, P. 91, 113, 141, 142,
 143, 209, 342, 409, 417, 420,
 421
 Clark, A. 409
 Cohen, M. 189, 202
 Cole, D. ix, 13, 42, 400, 405
 Collins, A. 125, 142, 143, 204
 Conant, J. 362
 Coolidge, C. 25
 Cooper, D. 413
 Copernicus, N. 119
 Cormann, J. 68
 Cotrell, G. 141, 142
 Cowan, J. 410
 Craik, F. 342, 417
 Creary, L. 405
 Cresswell, M. 415
 Crick, F. 166, 202
 Cullicover, P. 109, 409
 Cummins, R. 88, 200, 367, 372,
 373, 403
 Cussins, A. 141, 142

 Dahlbom, B. 67
 Dalai Lama 37
 Darwin, C. 271, 284
 Davidson, D. 9, 68, 69, 361, 396,
 398, 403, 415, 421
 Dawkins, R. 371, 373, 417
 De Sousa, R. 71
 Dehaene, S. 205
 Dell, G. 200, 202, 203
 Dellarosa, D. 200
 Demopoulos, W. 46
 Dennett, D. xi, 12, 68, 69, 71, 112,
 113, 142, 209, 248, 249, 250,
 332, 342, 367, 374, 404, 405,
 406, 410, 413, 415, 418, 423
 Derthick, M. 202, 410
 Descartes, R. 5, 31, 40, 54, 272,
 273, 283, 404
 Devitt, M. 361, 403, 410, 413

 Dilthey, W. 278, 280
 Dobzhansky, T. 373
 Donaldson, W. 144
 Donnellan, K. 257, 266, 347, 415
 Doyle, J. 339, 418
 Dretske, F. xi, 13, 15–18, 24, 25,
 27, 46, 113, 212, 213, 217, 218,
 219, 220, 221, 222, 223, 228,
 250, 413, 418
 Dreyfus, H. 19, 24, 91, 250, 406,
 410, 415
 Dummett, M. 403
 Dunlop, C. 342, 400
 Dykes, R. 422

 Eccles, J. 422
 Edelson, G. 411
 Edwards, P. 343
 Elvee, R. 280
 Erlbaum, L. 202, 203, 204
 Evans, G. 413, 418

 Franklin, B. 59
 Fanty, M. 200
 Farah, J. 410
 Feehan, T. 266
 Feigenbaum, E. 204
 Feigl, H. 54, 69, 144, 421
 Feldman, F. 69, 421
 Feldman, J. 146, 159, 196, 200, 201,
 202, 203, 204, 410, 412
 Fetzer, J. xi, 13, 43–46, 392, 393,
 394, 400, 401, 404, 421, 423,
 424
 Feyerabend, P. 69, 141, 143, 421
 Field, H. 210, 410, 413, 421
 Fingarette, H. 69
 Firth, R. 285, 302, 303
 Fiske, D. 344
 Flanagan, O. 403
 Fodor, J. xi, 6, 12, 13, 21–22, 24,
 25–29, 40, 41, 42, 43, 44, 54, 68,
 69, 88, 112, 118, 121, 128, 141,
 142, 143, 150, 151, 182, 183,
 200, 203, 205, 209, 210, 212,
 229, 230, 231, 232, 236, 239,
 241, 248, 249, 332, 343, 349,

- 358, 359, 362, 365, 366, 374,
 377, 379, 380, 381, 382, 386,
 387, 388, 389, 390, 391, 392,
 395, 396, 398, 400, 401, 404,
 406, 410, 412, 413, 418, 421,
 423
- Foerber, R. 42
- Foster, L. 68
- Foucault, M. 38, 289
- Freeman, W. 412
- Frege, G. 256, 305, 340
- French, P. 403, 418
- Freud, S. 6, 60, 67
- Galanter, E. 416
- Galileo, G. 119
- Garfield, J. 406, 414
- Garon, J. xi, 13, 21, 23
- Garrett, M. 413
- Gazzaniga, M. 421
- Geman, D. 203
- German, S. 190, 203
- Gettier, E. 11
- Geva, S. 412
- Gibson, J. 414
- Gilmartin, K. 342
- Glaser, R. 417
- Globus, G. 374, 421
- Glover, J. 421
- Godel, K. 296
- Goldfarb, R. 25
- Goldman, A. xi, 34-39, 287, 299,
 303, 309, 327, 341, 343, 414,
 418
- Goldstein, I. 307, 343, 418
- Goodman, N. 142, 143
- Grandy, R. 415
- Greene, H. 373
- Greeno, J. 410
- Gregg, L. 143
- Gregory, R. 404, 406
- Grice, H. 44, 66, 69, 416
- Grice, P. 82, 88n
- Gruber, H. 418
- Gunderson, K. 69, 361, 404, 406
- Gustafson, D. 69
- Guttenplan, S. 344, 404, 421
- Haldane, E. 404
- Hallam, A. 362
- Hamlyn, D. 69, 306, 343, 418
- Harman, G. 68, 69, 210, 320, 361,
 403, 414, 418
- Harnish, R. 405, 412
- Hartley, R. 249, 250
- Hatfield, G. 410
- Haugeland, J. 87, 88, 92, 118, 171,
 172, 203, 248, 250, 374, 387,
 401, 404, 406, 414, 415, 478
- Hayes, P. 184, 202
- Hebb, D. 159, 203
- Hegel, G. 283, 288
- Heidegger, M. 289
- Heil, J. 141, 249, 250
- Hempel, C. 250, 397, 401, 404, 421,
 424
- Hendrix, G. 407
- Higgenbotham, J. 414
- Hill, C. 416
- Hillis, D. 410
- Hills, D. 414
- Hintikka, J. 416, 418
- Hinton, G. 20, 146, 158, 159, 175,
 181, 187, 188, 190, 191, 196,
 200, 201, 202, 203, 204, 205
- Hirst, R. 418
- Hobbs, J. 406
- Hofstadter, D. 185, 203, 404, 410
- Holland, J. 142, 143
- Hollis, M. 418
- Holyoak, K. 143
- Hook, S. 70
- Hooker, C. 141, 143
- Hopcroft, J. 154, 203
- Hopfield, J. 410
- Horgan, T. 415
- Horwood, E. 205
- Hubel, D. 406
- Hull, C. 424
- Hume, T. 110, 298
- Jackson, F. 141, 142, 414
- Jakobovitz, L. 71
- James, W. 307, 404
- Jeffrey, R. 343, 419

- John, E. 372, 374
 Johnson, G. 410
 Johnson-Laird, P. 327, 343, 404,
 406, 420
 Jordan, M. 188, 203
 Joyce, J. 75
- Kalke, W. 69
 Kandel, R. 37, 374, 421
 Kant, I. 30, 110, 196, 203, 262, 283
 Kaplan, B. 340, 343, 419
 Kaplan, D. 69
 Katz, J. 69, 412, 414
 Kenny, A. 416
 Kim, J. 421
 Kintsch, W. 142, 143
 Kirsh, D. 141
 Kitcher, Pa xi, 12, 39–41, 141, 143
 Kitcher, Ph 141, 143, 144
 Klatzky, R. 334, 343, 419
 Klein, B. 69
 Kosslyn, S. 18, 92, 406, 410, 413
 Kripke, S. 31, 40–41, 59, 69, 70,
 347, 353, 354, 361, 410, 421,
 424
 Kuhn, T. 117, 141, 143, 343, 419
- Lambertian 110
 Larkin, J. 154, 203
 Lashley, K. 103, 159
 Lavoisier, A. 142
 Laymon, R. 406
 Leibniz, G. 6–7, 8, 55, 56, 70, 404
 Lepper 338, 340, 342
 Lesperance, Y. 407
 Levi, I. 424
 Levin, M. 361
 Lewis, D. 69, 156, 204, 361, 362,
 416, 421, 422
 Lindsay, P. 142, 144, 407
 Lloyd, D. 200
 Loar, B. 414, 419
 Locke, J. 404
 Lockhart, R. 342, 417
 Loftus, E. 316, 343, 419
 Lukes, S. 418
- Lycan, W. 69, 70, 141, 142, 144,
 372, 374, 404, 411, 416, 419,
 422
- Mach, E. 296
 Madell, G. 142, 144
 Malcolm, N. 50, 54, 70
 Maloney, C. 389, 401
 Margolis, J. 70
 Marks, C. 422
 Marr, D. 90, 407, 411
 Marras, A. 416
 Martindale, C. 404
 Matson, W. 70
 Maxwell, G. 69, 374
 McCarrell, N. 337, 343, 417
 McCarthy, J. 1, 46, 142, 144
 McCauley, R. 411
 McClelland, J. 20, 142, 143, 144,
 146, 147, 159, 162, 175, 178,
 188, 191, 196, 200, 201, 202,
 203, 204, 205
 McCulloch, W. 110
 McDermott, D. 378
 McDermott, J. 203
 McGinn, C. 404
 McIntyre, R. 416
 Medlin, D. 420
 Melden, A. 70
 Mellor, D. 424
 Mendel, G. 371
 Menzer, P. 266
 Mill, J. 36, 90, 288
 Miller, G. 327, 335, 343, 416, 419
 Minsky, M. 88, 204, 404
 Minsky, M. 70, 188, 200, 407, 411
 Mischel, T. 343, 419
 Moor, J. 407
 Moore, R. 283, 407
 Moravesik, J. 343, 419
 Morris, C. 379, 401
 Muntz, M. 69
- Nadel, L. 372, 374
 Nagel, E. 144, 364, 365, 366, 374
 Nagel, T. 63, 70, 141, 404, 416, 422
 Neisser, U. 404

- Nelson, R. 70, 416
 Newell, A. 103, 142, 144, 150, 151,
 204, 378, 401, 407, 411
 Newton, J. 111, 168
 Nickles, T. 143, 144, 374
 Nisbett, R. 143, 307, 343, 419, 422
 Norman, D. 142, 144, 404
 Norman, N. 407
 Nozick, R. 341, 404, 419
- O'Connor, J. 407, 422
 O'Keefe, J. 372, 374
 Otto, H. 405, 407, 416
 Owens, J. 422
- Palermo, D. 113, 342
 Palmer, S. 414
 Pap, A. 424
 Papert, S. 188, 204, 307, 343, 411,
 418
 Pappas, G. 303, 343, 419
 Parret, H. 422
 Parsons, K. 70
 Pearl, J. 190, 204
 Peirce, C. 44, 379, 392
 Perkins, M. 419
 Peters, R. 70, 416
 Piaget, J. 34, 307, 419
 Pinker, S. 92, 406
 Pitcher, G. 67, 70
 Place, U. 54, 70, 359, 422
 Plato 5, 44, 90, 212, 388, 389
 Poggio, T. 407
 Pollack, J. 200, 206, 407, 419
 Pollard, C. 405
 Popper, K. 393, 401, 422
 Posner, M. 327, 343, 419
 Powers, L. 333, 343, 419
 Presocratic atomists 19
 Pribram, K. 416
 Price, H. 52, 70
 Prior, E. 424
 Ptolemy 142
 Puccetti, R. 422
 Putnam, H. xi, 12, 36–39, 40, 41,
 54, 59, 69, 70, 71, 88, 209, 303,
 340, 343, 347, 352, 359, 361,
- 362, 364, 365, 366, 374, 385,
 386, 401, 405, 407, 408, 411,
 419, 424
 Pylyshyn, Z. xi, 18–19, 21, 24, 42,
 71, 97, 98, 111, 118, 128, 141,
 143, 150, 151, 183, 184, 200,
 203, 204, 205, 251, 332, 363,
 365, 368, 372, 374, 408, 410,
 411, 413, 414
- Quillian, M. 125, 142, 143, 144
 Quine, W. 34, 36, 38–39, 46, 52, 66,
 71, 209, 294, 295, 296, 297, 298,
 299, 300, 302, 303, 309, 343,
 344, 391, 398, 401, 420, 422,
 424
- Rankin, T. 400
 Ramsay, W. xi, 13, 21, 23, 142, 144
 Rapaport, W. 408
 Raphael, B. 408
 Rawls, J. 341, 344
 Reeke, G. 411
 Rey, G. 200, 411, 422
 Richardson, R. 365, 372, 374
 Riesbeck, C. 408
 Riley, M. 191, 192, 204, 205
 Ringle, M. 249, 251, 374, 408
 Rock, I. 414
 Rogers, A. 67
 Roller, D. 362
 Rorty, A. 71, 280
 Rorty, R. 38, 71, 289, 290, 301,
 302, 303, 405, 420, 422
 Rosenberg, C. 205
 Rosenschein, S. 406
 Rosenthal, D. 280, 362, 422
 Ross, G. 404
 Ross, L. 307, 338, 342, 343, 344,
 419
 Rozeboom, W. 424
 Rumelhart, D. 20, 142, 143, 144,
 146, 147, 158, 159, 162, 175,
 178, 188, 191, 196, 200, 201,
 202, 203, 204, 205, 411
 Russell, B. 64, 283
 Ryle, G. 11, 12, 49–50, 54, 60, 71,

- 209, 272, 280, 391, 401, 405,
424
- Salmon, N. 414
- Sampson, G. 411
- Savage, W. 69, 71, 88, 113, 373,
405
- Savodnik, I. 374
- Sayre, K. xi, 13, 26–29, 46, 249,
250, 251
- Schank, R. 8, 30, 201, 408, 411
- Schiffer, S. 405, 416
- Schmidt, F. 420
- Schneider, W. 412
- Scholes, R. 251
- Schwartz, S. 92, 406, 414
- Scriven, M. 69, 144
- Searle, J. xi, 8, 13, 15–18, 24, 27,
31–34, 46, 80, 88, 209, 251, 408,
416
- Segal, J. 417
- Sejnowski, T. 167, 181, 187, 190,
191, 200, 201, 202, 203, 205
- Selfridge, R. 249
- Sellars, W. 71, 142, 144, 251, 405,
416
- Shafer, R. 60, 71, 96
- Shaffer, J. 5, 51, 71, 405, 422
- Shaffner, K. 141, 144
- Shakespeare, W. 110
- Shannon, C. 249, 251
- Sharkey, N. 205
- Sharp, D. 410
- Sharpe, R. 142, 144
- Shastri, L. 190, 200, 205, 412
- Shepard, R. 163, 205, 414
- Sher, G. 361
- Shoemaker, S. 71, 349, 350, 405,
422
- Shope, R. 420
- Shweder, R. 344
- Simon, D. 203
- Simon, H. 142, 144, 150, 151, 203,
204, 341, 378, 407, 408
- Simon, T. 251
- Skarda, C. 412
- Skinner, B. 60, 62, 65, 68, 209, 391,
401, 405, 424
- Sloman, A. 408
- Smart, J. 54, 71, 273, 280, 359, 363,
364, 374, 422
- Smith, D. 416
- Smith, E. 420
- Smith, H. 341, 343
- Smith, N. 203
- Smith, G. 92, 406
- Smolensky, P. xi, 13, 21–23, 127,
128, 129, 142, 144, 157, 158,
175, 181, 184, 187, 188, 190,
191, 192, 195, 196, 197, 202,
204, 205, 412
- Soames, S. 414
- Sober, E. 372, 374
- Sosa, E. 420
- Sperber, D. 196
- Sperry, R. 67, 422
- Spiegelberg, H. 416
- Spiro, R. 204
- Stabler, E. 409, 412
- Stafford, S. 67n
- Stallman, R. 339, 344
- Stalnaker, R. 405, 415, 422, 424
- Stampe, D. 87, 212, 213, 214, 215,
216, 222, 223, 224, 228, 415
- Steinberg, D. 71
- Sterelny, K. 404, 413
- Stevenson, C. 424
- Stevenson, J. 273
- Stich, S. xi, 13, 21, 23, 45, 46, 72,
100, 121, 142, 144, 209, 379,
380, 386, 387, 401, 409, 412,
415, 417
- Strawson, P. 51, 71, 72, 303, 417
- Sussman, G. 339, 344
- Swain, M. 419, 420
- Swanson, J. 68
- Tarski, A. 297, 299, 300
- Taylor, B. 72
- Taylor, C. 72, 423
- Thagard, P. 143, 412
- Thatcher, R. 372, 374
- Tienson, J. 405, 415
- Titon, J. 67n

- Tolman, E. 62, 420, 424
Tormey, A. 72
Toulmin, S. 251
Toulouse, G. 201, 205
Touretzky, D. 184, 201, 205, 412
Troyer, J. 72
Tuedio, J. 405, 407, 416, 417
Tulving, E. 144
Tuomela, R. 424
Turing, A. 7, 58, 154, 161, 206, 409
Turvey, M. 415

Ullman, J. 154, 203, 420
Ullman, S. 105, 109, 409, 415

Van Gulick, R. 405, 409, 415, 417,
 423
Vemsom, J. 68, 72
Vendler, Z. 423
Von Eckardt, B. 423
Von Neumann, J. 4, 19, 138, 154,
 160, 161, 169, 172, 173, 176

Wade, M. 371, 374
Waldrop, M. 201, 206
Waltz, D. 200n, 206
Warner, R. 415
Warnock, G. 68, 72

Wason, P. 406, 420
Weaver, W. 249, 251
Weber, M. 98
Wegener, A. 358
Weimer, W. 113, 193
Weizenbaum, J. 409
Wexler, K. 107, 109, 409
Wheeler, S. 72
White, A. 72
White, S. 41, 423
Whitehead, A. 266
Wilkes, K. 142, 144, 423
Williams, B. 405
Williams, G. 371, 374
Williams, R. 158, 188, 201, 204
Wilson, D. 196, 371, 374
Wilson, T. 42
Wimsatt, W. 366, 371, 372, 374
Winch, P. 278, 380
Winograd, T. 98, 383, 401, 409
Winston, P. 204, 409
Wisdom, J. 72
Wittgenstein, L. 44, 50, 62, 67, 68,
 70, 72, 121, 286, 293, 390, 405,
 424
Wollheim, R. 68, 72
Woodfield, A. 417

INDEX OF SUBJECTS

- a conditional claim 118
a conditional thesis 137
A Materialist Theory of the Mind 11
a priori questions 5
ability to use a sign 400
about 57, 267
“aboutness” 4
absent qualia argument 6, 39, 41, 349ff; see also qualia
absent qualia examples 356
action/condition 174
activation evolution equation 157
activation levels 127
activation patterns 139
activation values 146
actual behavior 395
add and subtract 76, 78
address an utterance 262
addressing someone 262
after-images 56
agency 32
algorithms 238
ambiguity 382, 385, 390, 397
analog machines 14
analog parallel computation 150
analog/digital distinction 14
Analyse der Empfindungen 290
analytic epistemology 34
analytic/synthetic distinction 298
analytical epistemology 305
animals 45, 83–85
animals other than human 24
answers to questions 8
anti-realism 283–285
architecture 4
artifacts 1, 15, 33, 75–80, 222
artificial intelligence 1–2, 5, 7, 9, 16, 20, 67, 77–83, 147–148, 150, 156, 229ff, 307, 339, 377ff
artificial minds 15
ascriptions of intentionality 268–269
assertability 300f
assignment of the blame problem 182
associative memory 20, 24
asymptotic behavior 195
attribute 256
attribution of a property to something 29
automated formal systems 378, 387
automatic devices 15
autonomic nervous system 233
autonomous science 104
axiomatization 296
back-propagation procedure 201
backtracking 339
Bayesian 310
behavior (defined) 394
behaviorism 7, 11, 12, 29, 50–51, 60, 209, 230, 391, 396
behaviorists 124
belief 1, 4, 23, 30–35, 39, 84, 97, 121–126, 135, 141–142n, 209, 231ff, 256–266, 305ff, 392, 397
belief-change 340
beliefs and desires 4, 23
believing 256
Best Fit Principle 191
biological evolution 6
bivalence 296–297
blackboard model 339
Boltzmann machine 187, 191, 201f
brain 3, 170–171
brain function ‘A’ 233

- brain function 'C' 233
 brain state *C* 27–28, 242, 244–246
 brain states 17
 brain states embodying semantics
 27
 brains 286
 brittleness 12
 bundles of information 196
 by convention 386, 391
 by habituation 386, 391
 by nature 386, 390
- C-fibers firing 39–40, 353–354
 calculator 5, 78ff
 caloric theory 120
 canons of science 52
 capability to use a sign 400
 capacities 94
 capacity 286–287
 Cartesian dualism 54
 categories 197
 categorization 198
 category mistakes 49
 causal account of representation 212
 causal chains 17, 18
 causal conditions 222
 causal connections 28
 causal explanations 277
 causal properties 377
 causal relations 211
 causal theories 46
 causal theories of representation 218
 causal theory 43
 causal theory of reference 10, 40,
 46, 347, 350ff; see also rigid
 designator
 causality 10–11, 17–18, 25–28, 32,
 34, 43, 86, 121–122, 135,
 211, 212–227, 233, 261–265,
 247–280, 300, 370, 377
 causally active states 122, 140
 causation 216
 causes of beliefs 35
 central state materialism 347–348,
 350, 352–353, 359, 361
 cerebral cortex 165
ceteris paribus clauses 60
- change the sentence 215
 change the world 215
 channels 244–245
 Chinese 8
 Chinese Room 8
 Chinese symbol manipulation 80
 chunks 335
 Church's thesis 150, 161
 circuit state feature units 192
 classifications of contents 306
 clocks 14
 cognition 22
 Cognitive Science and the Problem
 of Semantics 26
 cognitive acts or processes 306
 cognitive capacities of machines 83
 cognitive competence 108
 cognitive contents 306
 cognitive faculties 306
 cognitive functions 168
 cognitive impenetrability 42, 97,
 369
 cognitive modeling 147
 cognitive models 126, 129
 cognitive operations 79
 cognitive organization 336
 cognitive processes 158, 168
 cognitive psychology 1, 9, 35, 65
 cognitive science 1, 2, 4, 7, 149,
 152, 198, 229f, 307ff
 cognitive significance in its broad
 sense 397
 cognitive significance in its narrow
 sense 396–397
 cognitive simulation 368
 cognitive systems 171, 179–183
 cognitive-state transitions 316
 combinatorial constituent structure
 23
 common sense mentalistic
 commitments 309
 common sense psychology
 120–121, 123, 139
 computational functionalism 233
 communication 7, 39
 communication theory 240
 communication-theoretic account

- 27–28, 229ff
competence 61, 174
competence of the system 193
competence/performance distinction 5, 193–195, 368
complex subsymbolic systems 168
compositionality 24
Computation and Cognition 18
computational conceptions 3–4, 9–10, 13, 44–46, 102–106, 149, 151, 209, 231ff, 377–378, 380, 382
computational model 231
computational neuroscience 149
computational power 148
computational processes 13
computational psychology 229
computational states 209
computational temperature 191
computational theories 3, 9, 43
computationalism 3–4; see computational conception
computers 1–2, 5, 15–18, 19, 26, 29, 75–87, 154, 160, 172, 229ff, 236, 365, m 377ff, 398ff
computer languages 378
computer models of cognition 3
computer program 1
computer programming 229
computer science 9
computer simulation 61
computer simulations 80
computers cannot add 16–17
computers cannot think 15
conception 85
concepts 5, 22, 54, 125, 309f, 393
conceptual analysis 30, 50
conceptual change 92
conceptual level 151, 156
conceptual role view 113
conceptual unit hypothesis 158
condition/action 174
condition of satisfaction 267, 275f
connection evolution equation 158
connection strengths 198
connectionism 2, 4, 19–25, 117–118, 130–134, 145–147, 126–202
connectionist computers 169
connectionist dynamical hypothesis 158
connectionist dynamical systems 165
connectionist hypotheses 126
connectionist memories 175
connectionist models 117, 120, 170
connectionist models as mere implementations 161
connectionist networks 130–134
conscious 45, 393
conscious concepts 159
conscious intentions 45
conscious rule application 17, 154
conscious rule interpretation 176
conscious rule interpreter 154, 170, 179
consciousness 13, 45, 75, 84, 90–92, 153–155, 170, 173–176, 234, 270–272, 396; see also qualia
consequentialism 313f
conservative theory change 130
constituent structures of mental states 9, 183–185; see also compositionality
constituents 151
constrained capacities 96
constructive psychological processes 316
content (or meaning) of specific beliefs 396
content 44, 61, 110, 229ff, 231, 259–264, 274–278, 381–384
content addressability 174
content of a concept 393–394
content of a representation 25
content of a sign 395
contents 56
contents of consciousness 175
context 385, 390, 393, 397
context-dependence 178–179, 185
continental drift 358
“Contingent Materialism” 42
continuity 186–189; see also digital/analog distinction

- control structures 103
 conveying meaning 261–266
 Copernican Revolution 21
 correspondence theory of truth 227,
 300
 counterfactual causes 219
 counterfactual situations 95
 counterfactuals 215, 217, 220,
 301–302, 395–396
 covering law 278; *see also* causality,
 explanation
 covering law explanation 367
 creativity 5–6
 “criterionless” 352
 criteria 53
 criterion, functional role 381
 criterion, parsing 381
 criterion, substitutional 381
 criticism 288
 crude approximations 170
 cultural activities 153
 cultural imperialism 39, 282, 294
 cultural norms 294
 cultural relativism 38, 287–292
- Darwinian evolutionary theory 371
 data structures 66, 112
 Davidson’s criticism of dispositions
 396
de dicto belief 259
de re belief 256–259
 “decision-guides” 314–315
 deductive logic 317
 deductive proofs 238
 definiendum 389
 definiens 389
 definition 389
 definitional circularity 390
 degrees of similarity of meaning
 396
 demonstratives 256, 264
 depth of processing 342
 derived intentionality 268–269
 describing the brain 169
 descriptions 102, 364
 descriptive AI 377
 descriptive epistemology 34, 305
- designation 263–265; *see also*:
 reference; causal theory of
 reference; *de re* beliefs
 desires 1, 275–277; *see also*:
 endeavoring; goal seeking;
 teleology
 devices 2
 dialects 398
 differential equations 195
 digital/analog distinction 14, 217
 digital computation 189
 digital computers 1–2, 7, 237
 digital forms 217
 digital machines 14
 direct belief 257
 discrete categories 186
 discrete computation 189
 discrete inference operations 186
 discrete learning operations 186
 discrete memory locations 186
 discrete memory operations 186
 discrete methods 188
 discrete production rules 186
 discreteness *see* continuity
 dispositional belief 333
 dispositional conceptions 10, 377
 dispositional properties 11–12
 dispositional states 139
 dispositional theory 43
 dispositions 11–12, 139–141, 230,
 284, 333, 391ff
 dispositions to behave 62
 dispositions toward behavior 395
 disquotation 301
 distributed systems 4
 Doctor Spock 76
 dogs 7, 16
 domain-independent rules 330
 doppelganger 236
 double-aspect theory 51
 doxastic states 312
 Dreaded Collateral Information
 Problem 214
 dualism 27, 31, 51, 54–55, 59
 Descartes’ argument for 40
 dynamical systems 158, 168, 187,
 189, 194–195

- elan vital* 55
 eliminationist line 299–301
 eliminative materialism 31
 eliminative theories 226
 eliminativism 12, 23, 31, 36, 46,
 117–142, 198–199, 209, 299
 emergent semantics 9
 emotions 1
 empirical boundary conditions 112
 empiricism 3
 endeavoring 30, 256, 260; *see also:*
 desire; teleology
 endeavoring to convey 261
 endeavors 261
 endeavors to bring about for the
 purpose of 261
 environmental state *E* 27–28
 environmental state *Ec* 28
 environmental states 27
 epiphenomenalism 54, 63, 270, 364;
 see also mind/body problem
 epistemic access theories 212–216
 epistemic consequentialism 313
 epistemic interdependence of
 motive and belief attributions
 397
 epistemic interdependence of
 motive, ..., and opportunity
 ascriptions 397
 epistemic possibility 40
 epistemic rule *R* 313
 epistemic rules 312, 319, 331
 epistemological constraints on
 semantics 9
 “Epistemology Naturalized” 52,
 294, 298
 epistemology 11, 13, 34–39, 52,
 305–340
 evolutionary 284–287, 307f
 genetic 307
 naturalistic 13, 283–287, 294,
 305–340
 normative 238, 310ff
 equivocation 244
 error-free learning 219
 eternal minds 5
 ethical consequentialism 313
 Evening Star 58
 events 6, 11, 56, 269f, 364
 evidence 299
 evidential indicators 53
 evolution 6, 19
 evolutionary epistemology 37,
 284–287, 307–308
 excitatory weights 146
 experiences 388
 explaining learned behavior 87
 explananda 18
 explanans 18
 explanation 18, 45, 85–112,
 118–120; *see also* covering
 law
 explanations and predictions 119
 explanations of behavior 33, 98
 explication 340
 extended traditional psychology 357
 extensionalistic 391
 “fact of the matter” 314
 false 256
 false belief 212
 falsehood 242
 finitely axiomatizable theories 296
 Fodor’s argument against disposi-
 tions 392
 Fodor’s theory of the language of
 thought 388
 “folk” psychology 4, 23, 117–142,
 230; *see also:* belief;
 eliminativism; propositional
 attitudes; consciousness
 folk theories 121
 for all I know 40
 form 44, 381–384
 formal grammars 5
 formal properties 380
 formal systems 378
 formality condition 26, 233ff, 341;
 see also computationalism
 foundationalism 38, 283, 298
 foundationalist epistemology 283
 framework for epistemology 318
 Fregean sense 256
From Folk Psychology to Cognitive

- Science* 46
 functional analysis 228, 358, 372
 functional architecture 42, 93, 105, 109, 362–363, 367ff, 372
 functional architecture of the mind 42
 functional constraints 015
 functional explanation 101
 functional level 3
 functional level of description 101
 functional role criterion 381, 398
 functional roles 58
 functional states 62, 209
 functional tendencies 46
 “Functionalism and Qualia” 349
 functionalism 11–13, 42, 44, 58, 60–61, 63, 100–106, 231, 272, 334, 347ff, 350–353, 360–363, 365, 372
 functionalist theory of mind 11
 functionalists 31
 functionally discrete 138
 functions 65
 fundamental epistemic rules 331
 fundamentally different kinds of systems 398
 fundamentally similar kinds of systems 399
 generality 326–327
 generalization 136
 generalizations 113
 generative grammar 10, 20
 genetic epistemology 307–308
 Gettier Problem 11
 ghost in the machine 249
 ghostly homunculi 250
 goal conditions 179, 201
 goal seeking 179–182, 397; *see also* teleology
 good approximations 170
 Grammar *G* 383
 grandfather clocks 14
 habits, skills, and dispositions 379, 389, 398
 hard constraints 194
 hard systems 23, 187
 hardware 2
 harmony theory 191–194
 hidden units 127, 146, 164
 high-level mental structures 198
 higher-level tasks 149
 higher-level theories 199
 historical reliabilism 309, 314
 historicism 288, 294
Homo sapiens 199
 homogeneous inhomogeneity 201
 homunculi 17, 249–250
 human actions 394
 human behavior 394
 human competence 149
 human language evolution 249
 human semiotic systems 394
 human thought about pi 57
 “Humpty-Dumpty” theory of meaning 10
 hyper-strong connectionism 24
 hyper-weak connectionism 19
 hypothesis, conceptual unit 158
 hypothesis, connectionist dynamical 158
 hypothesis, subconceptual level 159
 hypothesis, subsymbolic 160
 icons 392–393
 identities 57
 identity claims 59
 identity theories 43, 355
 identity theory 5–6, 12, 31, 39–43, 54–59, 230, 235, 272–274, 347ff, 363–364, 372
 idiolects 398
 idiot savants 76
 images 97, 111, 332
 implementations 160–161
 implementations of cognitive models 126
 in virtue of 214
 inborn primitive set of concepts 22
 inconsistent beliefs 36
 inconsistent information 194
 “incorrigibility” 12–13
 incorrigibility 67

- indeterminacy of radical translation 66
 indices 392–393
 indirect belief 257
 indirectly endeavors that 260
 individual knowledge 21
 individual mental events 41
 individual physical events 41
 individual units 159
 individuating functional states 103
 indivisible minds 5
 induction 288, 317; *see also:*
 inference; learning
 inexistence 64
 inference 4, 107, 123, 136–137,
 174–179, 182–183, 189–200,
 286, 313ff, 397, 398
 inference coherence 24
 inference procedures 183
 inferential network model 389, 397
 inferring 286
 infinite regress 390
 info(s) 239–241, 247
 info(s)-processing systems 241
 info(t) 27, 240–241, 244–245, 247
 info(t)-processing tasks 246, 248
 “information” 27
 information 16, 26–27, 42, 64–65,
 82, 97, 107, 213, 217, 228n,
 229ff, 239ff, 334
 information processing 19, 371
 information processing psychology
 91, 229
 information-theoretic sense 29
 inhibitory weights 146
 innate cognitive capacities 109
 innate representations 26
 inner processes 62
 input units 127, 146
 instrumentalism 60
 intelligence 1, 7, 15, 76–77, 229
 intensional properties 377
 intentional awareness 243
 intentional causation 32, 278–279
 intentional conceptions 10
 intentional events 268–269
 intentional explanations 275
 intentional idiom 230
 intentional instantiation 171
 intentional properties 210
 intentional sense of “goal” 180
 intentional states 32–33, 102, 209,
 268
 “Intentionality and its Place in
 Nature” 24, 31
 intentionality 4, 9, 29–34, 46,
 63–65, 102, 210, 227n, 233,
 243, 255–280
 intentionality of thought 255
 intentions 1, 4, 11, 26, 32, 222,
 260–263, 275–279, 358
 interactionistic dualism 54
 internal representations 66,
 235–236, 249
 interpretation 9, 13, 107, 174, 200
 interpretations 387
 interpreted formal systems 387
 interpreters 9
 interpreting mind 9
 interprets use of *P* with the
 attributive sense *S* 264
 intersubjectivity 295
 intrinsic intentional states 270
 intrinsic intentionality 31, 268–269,
 280
 intrinsic meaning 81
 introspection 6, 67, 110–112; *see*
 also self-awareness
 intuition 154ff, 190–191
 intuitive processors 155
 inverted spectrum 61–62; *see also*
 qualia
 involuntary 329
 IQ 76
 ISA relation 30
 justification 258, 299
 justified belief 36, 283, 311
 justified doxastic states 312, 329
 kind terms 230
 knowing how 390
 knowing that 390
Knowledge and the Flow of

- Information* 217
- knowledge 5, 11, 21, 52, 153–160, 177–179, 189–190, 198, 214
- knowledge atoms 192
- knowledge level 103
- knowledge of the system 127
- Kripke's theory 59
- language 1, 9, 29–30, 152, 198, 377
- language of thought 4, 17, 22 210, 332, 387ff, 390
- language processing 164
- languages 398
- lawlike regularities 126
- laws of physics 55
- learned symbols 219
- learning 83–87, 106–110, 163, 178f, 198, 201n, 219f, 286f, 387ff
- learning *L* 387–388
- learning procedures 127
- learning systems 157
- Leibniz' Law 55–56
- levels of analysis 150
- levels of description 10, 101
- linear systems 195
- linguistic behavior 7, 24
- linguistic competence 24, 174
- linguistic meaning 265
- linguistic primitives 391
- linguistics 1, 5, 7, 9, 18, 24, 61, 90, 162, 174, 178, 377, 398
- local connectionist models 195
- localist 126, 128
- logic 5, 35–36, 238, 317ff; see also: inference; induction
- logic and epistemology 36
- logical behaviorism 12, 50–54, 60, 62
- logical positivism 60
- Logische Aufbau* 290
- lower animals 45
- lower-level theories 199
- “Machines and the Mental” 15, 24
- machines 75
- macroinference 197
- manipulating systems 79
- mapping 232, 236
- marijuana 16
- massively parallel analog computers 149
- massively parallel distributed systems 4
- materialism 6, 10–12, 31, 41, 79, 272, 283, 289, 247ff; see also identity theory
- mathematics 5, 296f
- maximizing epistemic ends 38
- meaning (or content) of specific beliefs 396
- meaning 13, 30, 61, 64, 85–86, 234, 242, 265; see also content; intension; representation; sign
- meaning of a token 395
- meaning of symbols 78
- meaning to convey 260–261
- meaning to the machine itself 15
- meaningful behavior 99
- means 16
- mechanisms 12
- memory 130ff, 334ff; see also representation
- memory locations 334
- memory retrievability 336
- memory stores 334
- Mendelian genetics 371
- mental capacity 21
- mental classifications 359
- mental event types 364
- mental events 6
- mental lodging 335
- mental phenomena 1, 270
- mental properties 5
- mental representations 29, 309
- mental states 40, 183, 230, 365
- mental states as causal states 11
- mental terms 348
- mental tokens 381–382
- mentalese see language of thought
- mentality 4, 15, 45, 75, 377
- mere desires 32
- meta-epistemological principle 312
- meta-language 387

- metaphysics 283
metaphysical possibility 40
method 37
methodological solipsism 112, 232, 290–292, 297
microdecisions 192
Mills' Methods 36
mind 1, 7, 10
mind-independent facts 37
mind-talk 57
mind-to-world direction of causation 276, 279
mind/body problem 5, 12, 49, 272–274, 373; *see also*: behaviorism; dualism; epiphenomenalism; functionalism; identity theory; materialism
mind/brain identity theory 12, 39, 54
mindlessness 15
minds 5
minds and machines 67
minds of Type I 393
minds of Type II 393
minds of Type III 393
misinformation 217
misrepresentation 25–26, 212, 214–215, 217–218, 223, 225
missing qualia 6
mistakes 1, 201
modality-specific 332
modeling human performance 148
models for human belief 136
models for propositional memory 136
modularity, propositional 121–127, 135, 128–141
molecular motion 120
Monadology 7
Morning Star 58
Mueller-Lyer figures 42
multidimensional scaling 163
multiple-code theory of cognition 332
names *see* proper names
“Naming and Necessity” 40
narrow/wide psychological state 385f
nativist position 22
natural indicators 250
natural kinds 19, 89–93, 126, 136, 357
natural languages 230, 378
natural laws 94
natural meaning 82
natural sense of meaning 82
naturalistic accounts of intentions 26
naturalistic accounts of representation 25
naturalistic conditions for representation 210
naturalistic epistemology 36
naturalistic epistemology 38, 307–309; *see* epistemology, naturalistic
naturalistic theories in semantics 226
naturalized epistemology 13
nature/nurture distinction 42
negative feedback 33–34, 370; *see also*: goal; seeking; teleology
nested intentional systems 249
NETtalk 267–268
neural architecture 165
neural architecture hypothesis 157
neural level 164
neural models 167, 170
neuronal activity 10
neurons 165
neurophiles 209
neurophysiological explanation 98
neuroscience 1, 4f, 20, 42, 164–169, 235, 269, 360, 363ff, 369
neurotransmitter 17
neurophysiology 5
non-behavioristic 392, 398
non-connectionist cognitive models 128
non-digital transmission mechanisms 373
non-dispositional properties 12

- non-extensionalistic 392, 398
 non-linguistic understanding 389
 non-natural meaning 66
 non-physical events 55
 non-reductionistic 392, 398
 non-von Neumann architecture 373
 nonconnectionist paradigms 163
 nonlinear systems 195
 nonmonotonic inference 190
 nonmonotonicity 179
 nonrepresentation 225
 normal circumstances 221
 normal conditions 302
 normalcy conditions 25–26,
 221–225, 302n
 normative AI 377
 normative epistemology 34, 305,
 310
 object 259, 263
 observation sentences 295f, 302
 occurrence set 243
 occurrent belief 333
 “On the Proper Treatment of
 Connectionism” 21
 ontological inflation 9
 ontologically conservative theory
 change 120
 ontologically radical theory change
 120, 141
 ontology 9, 63, 120; *see also: anti-*
 realism; reductionism;
 identity theory
 operational definition 293
 operationalism 60
 operations 338
 optimal sets of rules 315
 ordinary language 51
 ordinary language analysis 12, 49,
 51
 ordinary language philosophy 49
 original intentionality 87
 other mental states 230
 other minds 53
 our conception of an epistemic rule
 329
 output units 127, 146
 P-knowledge 177–178
 pain 354, 360–362
 para-mechanical theoretical entities
 60
 parallel distributed processing
 (PDP) 201; *see also*
 connectionism
 parallel distributed processing 19;
 see also connectionism
 “parallelizing” serial algorithms 200
 parsing criterion 381, 383ff
 particulars 212
 pattern recognition 20, 83
 patterns of behavior 101
 PDP 201
 pegged observation sentences 302
 perceives 260
 perceiving 30, 256
Perception 52
 perception 6–7, 11, 16, 85, 90,
 242ff, 260–261
Perceptrons 188
 perceptually takes to be 260
 performance 61
 performance model 61
 performance of the system 193
 phenomenism 18
 phenomenalist accounts of
 dispositions 11
 phenomenological critiques 18
 phenomenology 18
 philosophical questions 293
 philosophy 1
 physical criteria 41
 physical event types 364
 physical events 5
 physical level 2
 physical level of description 101
 physical states 348
 physical symbol systems 378
 physical systems 5
 physicalism 230ff, 283, 289
 physicalists 31
 physiological theory 359
 physiology 52
 Plato’s theory of knowledge as
 recollection 388

- pocket calculators 5
 polyism 270
 positivism 38, 288, 298
 possibility, epistemic 40
 possibility, metaphysical 40
 practical reasoning 238
 pragmatic conceptions 44
 pragmatics 13, 44, 379
 prediction goal 181
 prediction-from-examples goal 181
 predispositions 400
 presuppositions of empiricism 3
 primitive set of concepts 22
 primitives 390
 private languages 398
 privileged access 12–13, 67
 problem of meaning (or content) 63
 problem of other minds 53
 problem space 379
 production system models 155
 productivity 24
 programming languages 33
 programs 127, 146
 projectable predicates 122, 125
 proof 106
 proof theoretic account of
 knowledge 182–183
 proper description of processing 147
 proper names 264; *see also: causal*
 theory of reference; natural
 kinds; reference
 property 256
 proposition 256
 propositional attitudes 64, 103,
 120–126, 136–137, 209–227,
 231ff, 233, 255–266; *see also*
 belief
 propositional content 26, 277–278
 propositional modularity 121–122,
 135, 138, 140–141
 propositions 29–30, 56, 64,
 124–125, 255
 psychoanalytic theory 60
 psycholinguistic theories 5
 psychological laws 364–365
 psychological models 129
 psychological primitives 391
 “psychological state” 356
 psychological states 40, 318
 psychological states in the broad
 sense 386
 psychological states in the narrow
 sense 386
 psychological theory 359
 psychologism 305
 psychology 5, 19, 35–36, 89–112,
 305–340, 347–372
 psychology of cognition 316
 psychophysical constraints 100
 PTC 147–148
 purposes 14
 Q-functions 350
 qualia 6, 13 39–40, 61–63, 250,
 347f, 360
 quasilinear systems 195
 questions and answers 7–8
R carries information about *S* 228
R represents *S* 210, 228
 radical theory change 130
 rational acceptability 283
 rational animal 199
 rational beliefs 37
 rationality 286, 298
 realism 12, 36–38; *see also anti-*
 realism
 realist accounts of dispositions 11
 realist notion of truth 36–38, 300
Reason, Truth, and History 39
 reasoning 111
 recollection, Plato’s theory of
 knowledges 388
 reduction 11
 reductionism 31, 36, 41, 365f, 371,
 378, 392
 reductionist behaviorists 209
 reductionistic 391
 reductionistic accounts of disposi-
 tions 11
 reductive theories 226
 reference 29–30, 151, 229, 231,
 234, 255, 263–266, 347; *see*
 also: causal theory of

- reference; intentionality; representation
- referents 10
- referring expressions 10
- reflexivity test 38
- relativism see anti-realism; cultural relativism; historicism
- reliabilism 37–38, 387, 309ff
- reliabilist account 37
- reliability 287
- reliable methods 287
- representation 14–15, 18, 23, 25, 28, 66, 78–79, 94–100, 127–128, 158, 162–164, 182–185, 209–227, 229ff, 242ff, 246, 267, 309ff, 379ff, 385ff; see also: belief; knowledge; propositional attitudes
- representation to world fit 34
- Representational Theory of the Mind (RTM) 386
- representational conceptions 4, 10, 25, 377, 385
- representational explanations 102
- representational states 102, 209
- representational subsymbolic system 22
- representational theory 43
- representationalism 4, 25–29, 44, 100–112, 209–227
- representations 18, 94, 96, 100, 107, 110, 127, 162–163, 380
- resemblance relations 211f, 223, 332; see also icons
- (RSG 1) 319–320, 326
- (RSG 2) 321–322, 326
- (RSG 2') 321
- (RSG 3) 324, 326
- (RSG 4) 324–326
- right-making characteristics 312
- rightly assertible 298
- robots 7, 16–17, 45, 65–66, 82–84
- RTM 386
- rule interpretation 175
- rule-following systems 22
- rules 154; see also: inference; linguistics
- rules and representations 20–22
- rules or principles 311
- S-knowledge 177–178
- schemata 196–198
- scientific adequacy of connectionist approach 147
- scope ambiguity 215
- scripts 8
- scripts and frames 196
- self-adaptive systems 108
- self-awareness 30, 84
- self-consciousness 30
- semantic ascent 50
- semantic content 229, 234–236, 246
- semantic information 229
- semantic level of description 101
- semantic networks 124–126, 129, 135
- semantic properties 122, 210, 232, 248
- semantic rules 27
- semantical properties 237
- “Semantics Wisconsin Style” 25
- semantics 4, 8–9, 18, 27, 29, 44, 101–103, 151, 158–159, 172, 182–185, 229ff, 255, 379ff; see also representation
- semiotic 379
- semiotic systems 393, 399
- semiotic systems of Type I 393
- semiotic systems of Type II 393
- semiotic systems of Type III 393
- sensations 1
- sense 255, 263–266
- Fregean 256; see also propositions
- sensors and effectors 16
- sentences 9
- sentience see qualia
- serial architecture 4
- Shakey 45
- sign processing capabilities 87
- signals 14
- “Signs and Minds” 45
- signs 392ff; see also representation
- simulated annealing 191

- simulation 7–8, 161, 189; *see also:*
 implementation; competence/performance
 distinction
- single-code theory of cognition 332
- singularity condition 25
- skill 390
- skilled performance 155
- smoke detectors 16
- social knowledge 21
- social scientific understanding 38
- soft constraints 189, 193–194
- soft systems 23, 187
- software 2
- solipsism 290; *see also* methodological solipsism
- solution space 379
- sources of knowledge 305
- speaker's meaning 261–265
- speakers 43
- speech acts 66
- spreading activation accounts 196
- standard usage 398
- state 390
- state vectors 157
- states 338
- statistical inference 190
- stereotypes 252, 352, 361
- stimulus meaning 295
- STM 380, 386
- strong conception of AI 399
- strong connectionism 20, 23–24
- strong thesis 317
- strong thesis of AI 378
- structured representations 24
- stupid homunculi 249
- sub-English-3 384–385
- sub-English-4 384–385
- subconceptual 159
- subconceptual level 151, 164, 166–170, 186; *see also* subsymbolic
- subconceptual level hypothesis 159
- subconceptual unit hypothesis 159
- subjunctive 302
- substitutional criterion 381
- subsymbolic 21–22, 126, 129, 150–152, 159–200
- subsymbolic architecture 165, 168
- subsymbolic computation 188, 194
- subsymbolic explanations 172
- subsymbolic hypothesis 160
- subsymbolic in relation to the symbolic 172
- subsymbolic inference 179
- subsymbolic models 166
- subsymbolic paradigm 150–152, 160–163, 167, 169, 172, 198
- subsymbolic processes 21
- subsymbolic rule interpreters 176
- subsymbolic research 199
- subsymbolic semantics 182
- subsymbolic system 22
- subsymbolic systems 173, 190
- subsymbols 151
- success verb 267
- successor theories 119
- supernaturalism 31
- symbol manipulating engine 7
- symbol manipulation 80–81, 88
- symbol system 378ff
- symbol system hypothesis 378
- symbol users 45
- symbolic 171
- symbolic activities 79
- symbolic explanations 172
- symbolic paradigm 150–151, 160–161, 186, 200
- symbols 13, 22, 83, 129, 156, 393
- synonymy 385, 397
- Syntactic Theory of the Mind (STM) 380, 385f; *see also* computationalism
- syntactic engines 5
- syntactic properties of a symbol 13
- syntactic types 13
- syntax 7–9, 13–15, 161, 232, 237, 378ff
- system of knowledge 297
- “system of the world” 296
- systematic explanation 171
- systematic reduction 171
- systematicity 24

- Tarski's definition of truth 297, 299
 technology 2
 teleological forms of explanation 278
 teleological sense of "goal" 180
 teleology 221–226, 278f; *see also:*
 goal seeking; negative feedback
 tendencies to behave 46
 "The Appeal of Parallel Distributed Processing" 20
The Concept of Mind 11–12, 49
 "The Constituent Structure of Connectionist Mental States" 22
 "The Functional Architecture of the Mind" 41
The Language of Thought 44, 239; *see also:* mentalese; language of thought
 "The Primacy of the Intentional" 29
 "The Relation between Epistemology and Psychology" 34
 "The Robot Reply" 16
The Structure of Scientific Revolutions 117
 the missing qualia objection 6
 the physical and the mental 6
 theorem-proving 16; *see also:* inference; logic
 theorem-proving programs 238
 theoretical connectionism 19
 theoretical entities 60
 theories of mind 49
 theory change 120
 theory construction 50
 theory of cognition 393
 theory of reference 10
 theory of signs 379
 thermometers 222
 thermostats 14f
 things 51, 110–111
 thinking 286
 thinking beings 75
 thought 29–30, 57, 255
 thought about 57
 thought about pi 57
 thought processes 5
 thoughts 56, 380, 400
 three cognitive systems 171
 token identity theory 58
 tokening 217–220, 223
 token-physicalism 231, 235, 238
 token-token identity theory 364
 tokens 44
 tokens and types 381
 top-down 229, 233
 top-down causation 370
Tractatus 293
 translation 16
 tree rings 223–224
 true (rightly assertible) 293
 true belief 35–37
 truth 9, 212, 226, 231, 234, 238, 242, 284ff, 286–287, 300–301, 303, 387f, 394; *see also:* anti-realism; misrepresentation; representation
 truth and reference 232
 truth definition 297
 truth of a goal state description 34
 truth preserving rules 27
 truths 296
 Turing machines 2, 58–59, 148, 161
 Turing Test 7–8, 15
 Twin Earth 218
 type identity theory 58
 type-type identity theory 364–365; *see also:* identity theory
 type-type theories 355
 type/token ambiguities 211
 type/type identity 41
 uncertainty 27
 unconscious rule interpretation hypothesis 156
 understanding 1, 7–9, 16, 75, 77–87, 291
 understanding language 8
 uninterpreted formal systems 387
 unit's value 158, 195
 units 146
 Universal Turing Machines 148

- unlearned representation systems 25
uses *P* with the attributive sense *S*
 263
uttering 262–266, 291, 311, 352
- veridicality 182, 201
verificationism 53, 62, 226–227; see
 also: positivism; logical
 positivism
- virtual machines 21, 153–154, 170,
 173, 368
- visual perception 90
- visual thinking 91
- volition 276
- voluntary 329
- von Neumann architecture 4, 19
- von Neumann computer 154, 169;
 see also: computer
- von Neumann machines 161, 170,
 173; see also: computer
- warranted assertability 27, 238
washing clothes 337
weak conception of AI 399
weak connectionism 20–21, 23
weak thesis 317
weak thesis of AI 378
weights 127
“What Psychological States are
 Not” 40
- what a symbol means 216
what a symbol represents 216
- what does a dog believe? 84
- what is represented 96
- “What’s In a Mind?” 18
- “Why Reason Can’t be Naturalized”
 36
- widely distributed 126, 128
- wild tokenings 218
- words 51
- world-to-mind direction of fit 276,
 279