

---

# THE PHILOSOPHY OF DANIEL DENNETT

---

EDITED BY BRYCE HUEBNER

# THE PHILOSOPHY OF DANIEL DENNETT



# THE PHILOSOPHY OF DANIEL DENNETT

---

*Edited by* Bryce Huebner

OXFORD  
UNIVERSITY PRESS

**OXFORD**  
UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford. It furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide. Oxford is a registered trade mark of Oxford University Press in the UK and in certain other countries.

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

© Oxford University Press 2018

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

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

CIP data is on file at the Library of Congress  
ISBN 978-0-19-936751-1

9 8 7 6 5 4 3 2 1

Printed by Sheridan Books, Inc., United States of America

# CONTENTS

<i>Acknowledgments</i>	vii
<i>List of Contributors</i>	ix
<i>Introduction: Dennettian Themes Will Not Go Away</i> —BRYCE HUEBNER	xi

## PART I *Person-Level and Subpersonal Explanations*

1.1 Embodied Stances: Realism Without Literalism—REBECCA KUKLA	3
1.2 Reflections on Rebecca Kukla—DANIEL C. DENNETT	32
2.1 The Many Roles of the Intentional Stance—TADEUSZ ZAWIDZKI	36
2.2 Reflections on Tadeusz Zawidzki—DANIEL C. DENNETT	57
3.1 Memory and the Intentional Stance—FELIPE DE BRIGARD	62
3.2 Reflections on Felipe De Brigard—DANIEL C. DENNETT	92
4.1 Representations and Rules in Language—RAY JACKENDOFF	95
4.2 Reflections on Ray Jackendoff—DANIEL C. DENNETT	127

## PART II *Conscious Experience*

5.1 Seeming to Seem—DAVID ROSENTHAL	133
5.2 Reflections on David Rosenthal—DANIEL C. DENNETT	165
6.1 Is Consciousness a Trick or a Treat?—JESSE PRINZ	171
6.2 Reflections on Jesse Prinz—DANIEL C. DENNETT	196

7.1 Strange Inversions: Prediction and the Explanation of Conscious Experience—ANDY CLARK	202
7.2 Reflections on Andy Clark—DANIEL C. DENNETT	219
PART III <i>Evolution, Sociality, and Agency</i>	
8.1 Towers and Trees in Cognitive Evolution—PETER GODFREY-SMITH	225
8.2 Reflections on Peter Godfrey-Smith—DANIEL C. DENNETT	250
9.1 Mother Culture, Meet Mother Nature—LUC FAUCHER AND PIERRE POIRIER	254
9.2 Reflections on Luc Faucher and Pierre Poirier—DANIEL C. DENNETT	290
10.1 Planning and Prefigurative Politics: The Nature of Freedom and the Possibility of Control—BRYCE HUEBNER	295
10.2 Reflections on Bryce Huebner—DANIEL C. DENNETT	328
11.1 Dennett on <i>Breaking the Spell</i> —LYNNE RUDDER BAKER	331
11.2 Reflections on Lynne Rudder Baker—DANIEL C. DENNETT	345
<i>Index</i>	355

## ACKNOWLEDGMENTS

This anthology has been in the works for quite a while, and I have accumulated a great many debts in the course of editing it. First, I would like to thank Peter Ohlin at Oxford University Press for taking on this project, and for nudging it along when it seemed to be running into roadblocks. Second, I would like to thank all the amazing people who contributed to this volume. The level of engagement with Dennett's work is astounding, and I have learned a great deal from reading these essays! I hope that others will learn as much—as I think that each of the contributions opens up numerous avenues for exploring the terrain of Darwinian humanism, my preferred term for Dennett's theoretical enterprise (the section on consciousness is missing one planned article; the world conspired against its completion, and I flag it only because it feels to me like a hole). I would also like to thank everyone who served as external referees for the papers in this volume; this includes Cameron Buckner, Ruth Kramer, Rebecca Kukla, Pete Mandik, Amber Ross, Carl Sachs, and Tad Zawidzki; Felipe De Brigard deserves a special acknowledgment in this regard, as he went above and beyond the call of duty, reading multiple papers and offering deeply insightful comments on things that I would have missed. I hope that I haven't forgotten anyone.

Finally, and most importantly, I would like to thank Dan Dennett. His reflections here are characteristically insightful, and some of them open up exciting and fruitful avenues for future investigations. Many of them point to new ideas that he has developed in his new book, *From Bacteria to Bach and Back*. Having read a draft of the manuscript, I can confirm that Dennett is pushing his project forward and building on exactly the kinds of insights that his friends and critics have been pushing him on. Dan has been an amazing mentor to me over the years. He has introduced me to an incredible number of new ideas, supported me intellectually and as a friend, and he has taught me about the importance of criticizing with kindness. Just as importantly, Dennett has helped me to see that philosophers can look ahead, chart out new territories, and uncover previously unacknowledged possibilities. This is the deeply Darwinian insight that lies at the heart of the philosophical project; and it's also the reason why Dennettian themes will not go away.





## CONTRIBUTORS

**Lynne Rudder Baker**

University of Massachusetts at  
Amherst.

**Felipe De Brigard**

Duke University.

**Andy Clark**

University of Edinburgh.

**Daniel C. Dennett**

Tufts University.

**Luc Faucher**

Université du Québec à Montréal.

**Peter Godfrey-Smith**

University of Sydney and Graduate  
Center, CUNY.

**Bryce Huebner**

Georgetown University.

**Ray Jackendoff**

Tufts University.

**Rebecca Kukla**

Georgetown University.

**Pierre Poirier**

Université du Québec à Montréal.

**Jesse Prinz**

Graduate Center, CUNY.

**David Rosenthal**

Graduate Center, CUNY.

**Tadeusz Zawidzki**

George Washington University.



## INTRODUCTION: DENNETTIAN THEMES WILL NOT GO AWAY

Bryce Huebner

Daniel C. Dennett began publishing innovative philosophical research in the late 1960s, and he has continued doing so for the past 50 years. Over the course of his career, he has defended a robust form of Darwinian naturalism, and a passionate brand of humanism; and he has tackled a cluster of “big questions” about the nature of intelligence, agency, consciousness, and culture. Dennett has also worked to make his views accessible to a broad range of audiences—both inside and outside philosophy—by avoiding disciplinary jargon and using playful examples to show how biological, psychological, and social facts hang together. But Dennett doesn’t want to tell readers what to think—he wants to provide them with better tools for thinking *about thinking*, and he wants to show them that some of their treasured tools for thinking aren’t quite as useful as they seem.

The contributions to this volume examine the boundaries of Dennett’s research program, and clarify the conceptual resources Dennett offers for understanding mentality, evolution, and sociality. The four chapters in Part I focus on Dennett’s account of person-level and subpersonal explanations; two of these chapters target the nature of the intentional stance (Kukla, chapter 1.1; Zawidzki, chapter 2.1) and two examine core aspects of subpersonal cognitive psychology (De Brigard, chapter 3.1; Jackendoff, chapter 4.1). The three chapters in Part II address Dennett’s account of conscious experience (Rosenthal, Prinz, Clark, chapters 5.1, 6.1, and 7.1, respectively). Finally, the four chapters in Part III examine Dennett’s particular form of gradualism about biological evolution (Godfrey-Smith, chapter 8.1), his memetic account of social evolution (Faucher and Poirier, chapter 9.1), his views about human freedom (Huebner, chapter 10.1), and his claims about religious beliefs (Baker, chapter 11.1). My aim in this introduction is to provide an overview of some of these themes, and to clarify some of the critical insights they raise about Darwinian humanism. But I’ll start with the intentional

stance, as it provides the core of Dennett's approach to mind, evolution, and sociality.

## 1. The Intentional Stance

Few people have explicit commitments about what it means for gnats, bats, rats, cats, people, and corporations to act intentionally. But most form assumptions about the beliefs, desires, hopes, and wishes that these entities rely on to act intentionally (Wegner & Gray, 2016). In describing what it means to adopt the intentional stance, Dennett (1987) argues that such hunches arise through a simple inferential process:

First you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in most instances yield a decision about what the agent ought to do; that is what you predict the agent will do. (p. 61)

According to Dennett, we attribute the beliefs and desires that an agent *ought* to have whenever we want to predict, explain, or understand its behavior, and we interpret its behavior in light of these rationally structured attributions. In doing so, we are adopting an intentional stance.

Consider a disagreement between two friends. Femi claims that jollof rice is a Nigerian dish; Kofi replies that it's a Ghanaian dish; and each of them argues that it matters who's right. To understand this dispute, we don't need to know what happens in their brains; but we do need to assume that Femi and Kofi believe what they say about the origin of jollof rice. And we need to treat them as rational agents, who have the beliefs and desires they should have, given their location in material, cultural, and social space. But understanding this dispute is not a matter of uncovering beliefs that lie behind the statements that occur in their conversations. It requires making assumptions about what would make it rational for each of them to behave as they do, and situating their claims within larger patterns of rationally intelligible behavior. Of course, we typically assume that others have rational commitments that are roughly like our own; and this provides an initial foundation for our interpretive practices. But rational action often depends on the specific values and interests an individual has taken on; and rational actions are often shaped by particular learning histories, as well as unique

and idiosyncratic commitments. So successfully adopting the intentional stance often requires knowledge of normative commitments, and knowledge of material and social situations. So learning about Femi's cultural ties to Nigeria, and about Kofi's cultural ties to Ghana, can help us understand the nature of their disagreement; and learning about the cultural value of jollof rice can help us to understand why they care about the origin of this dish.

These claims probably seem banal. And they should. For many of us, the adoption of the intentional stance is reflexive and automatic. We treat the baristas at our favorite cafés as rational agents, and we assume that they will make excellent coffee—because that's what baristas *should* do. But few of us are mindful of drawing inferences about their behavior, and few of us make time to think about the beliefs and desires that they *should* have. For many of us, explicit inferential reasoning only occurs if we face a communicative breakdown. In part, this is because neurotypical people are disposed to assume that agents will act to achieve their goals in the most efficient ways available to them (Gergely & Csibra, 2003). These dispositions provide the foundation for the folk-psychological practices that guide our attempts to predict, explain, and understand one another as agents. Through repeated encounters with the world, we learn that certain entities will tend to possess locally-salient goals, and this structures habitual thought patterns about what those entities believe and desire. And by holding one another to shared norms of rationality, and teaching and learning from one another, we become more readily interpretable to one another as agents (McGeer, 2007; Zawidzki, 2013). Through this process, the mental states that we attribute from the intentional stance become *conventionally real* constituents of our shared world.

The upshot is that the mental states of folk psychology are not self-standing entities. They are theoretical projections that we paint onto the world through the collaborative practices of giving and asking for reasons. So the existence and character of mental states always depends on interpersonal practices of mindshaping, and ongoing engagements with the material and social world that we inhabit. As the thought patterns that guide this process fade into the background, and as we make one another more predictable through practices of mindshaping, our ascriptions of mentality become precise, accurate, and automatic. We only need to re-examine our expectations if we fail to predict another agent's behavior (Friston & Frith, 2015). This fact has obvious payoffs when it comes to social fluency; but it also has serious drawbacks. Many of us assume that in-group members have mental lives that are richer than the mental lives of out-group members (Spaulding, 2017); so we tend to draw problematic inferences in interpreting the behavior of people who differ from us in significant ways. Moreover, many of us find it difficult to understand forms of neurodivergent agency that are not

predicted by our learned expectations (Huebner & Yergeau, 2017). Together, these facts about the intentional stance help to make it clear why it cannot be a disinterested theoretical standpoint: our ascriptions of mentality are always grounded in learned and evolved expectations about who counts as an agent, and what it is rational for them to do.

Rebecca Kukla pushes this insight even further in her contribution. Against the common assumption that stances are mindsets, which we can adopt at will, she argues that stances should be understood as “collections of *embodied, performed strategies for coping and coordinating* with the world and the people in it” (p. 8). When we prepare to cope with some aspect of the world, our attention shifts, distinctive patterns show up to us, and we reflexively form expectations about what’s likely to happen next. In this process, some patterns prove more resilient and central to our coping strategies than others. For example, beliefs, desires, and reasons show up to us as indispensable posits when we engage with other as agents. Finally, she argues that there’s no intelligible way to ask whether mental states are real simpliciter, as opposed to real to someone who adopts the intentional stance. To make a claim about reality simpliciter, we would need to adopt a stance-free perspective; but this is impossible, as every engagement with the world has an embodied and performative dimension, and every engagement with the world evokes patterns of attending and expecting that can’t be de-stanced.

But is Dennett right that states like beliefs, desires, and intentions are only intelligible from inside the intentional stance? In his contribution, Tadeusz Zawidzki argues that the intentional stance can’t play all the roles Dennett has typically claimed that it must play. Zawidzki is skeptical about the suitability of the intentional stance as an analysis of mature, person-level, intentional concepts, in part because of the dynamic and socially situated structure of our interpersonal practices. He argues that folk ascriptions of mentality are often guided by ongoing regulative concerns with impression management and identity construction. Nonetheless, he argues that scientific practice often relies on intentional states that are characterized in terms of their predictive and explanatory roles; and he argues that most humans employ tacit cognitive resources with a similar character when they make quick and efficient behavioral anticipations. So he agrees with Dennett that the intentional stance provides a viable “model of quotidian interpretation, an account of the nature of the mind and mental states, a strategy for naturalizing the mental, and an important component of sound methodology in cognitive science” (p. 37). But he suggests that it is unlikely that a single set of explanatory norms will be operative in practices of quotidian interpretation, scientific explanation, and philosophical naturalization.

## 2. The Personal-Subpersonal Distinction

When we use the intentional stance to understand ourselves, or to understand one another, we often discover ways of using reason-based interventions to shape behavior. We can ask people to do things, and their responses frequently reveal an understanding of our request *as a request*. And we can ask ourselves to do things, and often, we seem to comply (unless we are tired, hungry, anxious, or otherwise undermotivated). But we shouldn't expect any particular response when we ask a cat or a rat to do something, and our reason-giving practices are useless when it comes to interacting with nonhuman animals (Roepstorff & Frith, 2004). Without linguistic competence, cats, rats, bats, and other animals cannot understand reasons *as reasons*. But this doesn't mean that they are not intentional systems. Even so, our explanations of their behavior do appeal to their reasons for acting. And these explanations are not structured interpersonally; they arise from within a predictive and explanatory perspective that doesn't assume that they understand why they act as they do. Put differently, we treat other animals as systems with purposes designed by natural selection, as well as by adaptive capacities they have acquired through learning. And according to Dennett, understanding these evolutionary and learned processes is essential to an adequate science of the mind, even if it requires us to take up a slightly different perspective than we adopt in engaging with participant in shared and cooperative activities.

Just as importantly, many of our practical capacities are uneasy targets for person-level explanations. Like all other animals, we are creatures of habit, and many of our habits are shaped by simple forms of learning. Behaviorist training strategies work well for getting rats to run mazes, cats to eat at particular times, and people to respond to motivationally salient phenomena. And when human or nonhuman animals act on learned associations, practices of reason giving and reason taking tend to be less effective in shaping behavior. Yet actions driven by learned associations are rationally intelligible, and relying on assumptions about what agents should believe and want, given the kinds of entities they are, allows us to work out what it would be rational for them to believe, and what it would be rational for them to do. Indeed, figuring out successful conditioning strategies *requires* treating an animal as an agent, and using the intentional stance to discern real patterns in their behavior. This use of the intentional stance moves us outside of the role of participants in shared practices, and into the role of observers of ongoing behavior. But to avoid the Cartesian assumption that minds are *places* where mental states occur as self-standing entities, Dennett also needs to deny the common assumption that the taxonomy of folk psychology provides the most appropriate subject matter for a biologically grounded and mechanistic theory of behavior (cf., Akins, 1996, p. 339). This is why Dennett draws the distinction



between person-level explanations and subpersonal explanations: the different idealizing assumptions make it unlikely that we'll find any neat mappings between the categories of folk psychology and the categories required to understand how brains encode information that's useful for their purposes; moreover, the predictive success of the intentional stance at the person-level does not guarantee that we are tracking a real pattern at the subpersonal level, nor does the predictive success of the intentional stance at the level of subpersonal processes guarantee the presence of a corresponding real pattern at the person level.<sup>1</sup> So how should we understand the relationships between these two explanatory projects?

There are two ways that philosophers tend to interpret this relation (Drayson, 2014). Some suppose that the person-level states are *sui generis* states and that any attempt to explain them by appealing to neural mechanisms will be incoherent. Others reject Dennett's conventional realism and assume that beliefs and desires are subpersonal states that can be characterized without recourse to the interpretive norms of the intentional stance; for them, beliefs and desires are person-level states, computational states, and states of the brain. This yields continuity between person-level states and subpersonal states, but this requires positing internal states with robust semantic content, and retreating from the kind of social ontology that Dennett prefers (and thus threatens a retreat to the Cartesian assumption that the mind is a place, populated by things). Dennett attempts to steer a middle course between these positions. He argues that person-level explanations always operate horizontally, appealing to person-level states as the cause of person-level behavior; and he claims that subpersonal explanations aim to demonstrate causal links between the operation of subpersonal capacities and the system-level behavior that they necessitate. Just as importantly, he argues that we should characterize subpersonal capacities intentionally (yielding continuity between person-level and subpersonal states), "our brains—more particularly, parts of our brains—engage in processes that are strikingly like guessing, deciding, believing, jumping to conclusions, etc." (Dennett, 2007, p. 86). And since the behavior of these systems is similar enough to the behavior we observe in other agents (people, cats, bats, rats, or gnats), we extend the intentional stance to them as well.

This means that persons are constituted by subpersonal agents, and that subpersonal explanations can be continuous with person-level explanations because they both operate from the perspective of the intentional stance (Drayson, 2012). But we will often find that the ends that drive subpersonal processes diverge from, and even conflict with, person-level goals and interests. From the perspective of

---

1. Note: Thanks to Felipe De Brigard for pushing me to clarify this point.

subpersonal cognitive psychology, the human mind consists of a wide variety of selfish computations (Huang & Bargh, 2014), narcissistic processes (Akins, 1996), and coalitions of cooperating systems (Huebner & Rupert, 2014), all of which interact to collectively guide behavior. In his most recent work, Dennett has argued that neurons behave selfishly, and he argues that we should adopt the intentional stance toward them as well. Following Turing (1950), he argues that we should repeat the process of functional and intentional decomposition until we reach a level where systems can be characterized purely mechanically. But he contends that this is only likely to happen at the level of motor proteins and their ilk.

This is what it means to say that we would adopt the intentional stance *almost all the way down*). This means that explanation must flow in two directions: upward from simpler to more complex mechanisms; and downward from more complex intentional capacities to simpler ones. And since all subpersonal explanations take place within the space of reasons, we should not treat person-level states as *sui generis*, even if subpersonal states and processes do have their own standards of rationality. Working back upward, Dennett aims to show that mechanisms are operative at every level of explanation. And strikingly, this means that even person-level transfers of information can be characterized mechanically.<sup>2</sup> At every level of explanation, mechanical processes shape the operation of subpersonal processes as cascades of information flow downward through the hierarchically organized network of computational processes (cf., Clark, this volume). Dennett thus contends that we should treat person-level content and person-level consciousness as arising from the mechanical operation of the brain, which continuously integrates numerous sensory inputs with a complex network of intentionally characterized “top-down” predictions; and he holds that person-level experience should thus be characterized in terms of the interpersonally constructed, and interpersonally accessible content that is available to agents who represent themselves and who represent one another as entities with reasons (cf., Shea et al. 2014).

---

2. There are at least two ways of making this kind of point. First, we might follow Bill Lycan (1990) in construing the successive levels of homuncular decomposition realistically, and assuming that all of our person-level explanations point toward a stable hierarchy of computational mechanisms. If this is right, we should expect to find parallel decomposition arising from the perspective of the physical, design, and intentional stance. Alternatively, we might follow Jay Rosenberg (1994) in treating the successive levels of homuncular decomposition as ontological fictions, construing explanation realistically only at the beginning and end of the story. I see compelling arguments in favor of both approaches, and I'm not sure which approach Dennett prefers; nor am I sure which approach makes the most sense of scientific practice. Consequently, I've tried to leave both options open in the main text. Thanks to Felipe De Brigard for asking me to clarify these points.

In his contribution to this volume, Felipe De Brigard builds on this perspective and argues that research from cognitive and computational neuroscience supports a Dennettian approach to episodic memory. He reviews evidence from cognitive psychology, neuroscience, and neuropsychology to defend a functional thesis about how memory works—he argues that memory systems process “information about past events, and use this information to construct useful anticipations of possible future events” (p. 64). De Brigard then reviews work in computational psychology and cognitive science to support a subpersonal account of the computations that are required for episodic remembering; here, he argues that statistical regularities, along with an individual’s limitations and goals, probabilistically constrain the search-space that will be examined during memory retrieval. Finally, he concludes by defending an ontological hypothesis about episodic memory as a person-level process, which is best understood from the intentional stance and not from a view that focuses exclusively on subpersonal structures that encode intentional contents.

Perhaps more than any other chapter in this volume, De Brigard’s contribution makes it clear how the intentional stance can be enhanced by close attention to subpersonal cognitive psychology. But he also shows that examining a person-level capacity from a subpersonal perspective can lead us to uncover a network of constituent processes, each of which can be characterized in intentional terms. These are the kinds of issues that Ray Jackendoff takes up in his contribution, as well. He addresses the conceptual difficulties that arise in attempts at reverse engineering the capacity for language, and he begins by noting that subpersonal explanations require characterizing three distinct types of phenomena:

1. the data structures that are employed in a domain (e.g., words or morphemes);
2. the computational processes that operate in that domain (e.g., the rules that allow us to build sentences out of words); and,
3. the interfaces that link domains to one another (e.g., syntax, semantics, and phonology).

Alternative hypotheses about the rules and representations that are operative in linguistic competence are rarely compared in detail. In part, this is because most people are trained in a single tradition, and they rarely consider the costs and benefits of continuing to work in that tradition. But Jackendoff provides a detailed comparison of derivational, nonderivational, and rule-free approaches to linguistic competence. Over the course of his chapter, Jackendoff provides tentative support for a nonderivational approach to the language faculty, which treats rules and representations as linked data structures that encode phonological,

semantic, and syntactic information. This allows him to see rules and representations as varying continuously along two dimensions: from fully specified words to unspecified variables, and from partially productive rules to fully productive rules. Whether his approach to linguistic competence is right or not, his paper provides a valuable attempt to examine where rival research programs stand on the nature of rules and representations; and it provides a useful resource for anyone who wants to consider the status of the rules and representations that are at play in other cognitive domains.

### 3. Subjective Experience

Summarizing, we might say that Dennett defends a form of interpretivism about person-level mental states, and an empirically driven—though deeply normative—approach to subpersonal and mechanistic explanation. Many philosophers grant that this approach illuminates some aspects of mentality, but apprehensions set in when we turn to questions about consciousness, freedom, and agency. Specifically, people worry about (a) Dennett's willingness to downplay the significance of first-personal data for answering hard questions about consciousness (e.g., Chalmers, 1996); (b) his willingness to treat consciousness as an illusion; and (c) his willingness to pursue questions about agency within a mechanistic and evolutionary framework. I address worries about consciousness in this section, and questions about freedom and agency in the sequel.

Bartenders can mistake a cider for a beer, but they cannot make a similar mistake about how the cider tastes to them, even if it tastes like a beer—or so it is often supposed. When I say that a cider tastes sweet, I report on my experience, and the truth of my report is rarely a matter for further investigation. And when I feel angry or sad, I seem to be the best judge of how I feel, even if I'm not always the best judge of why I feel as I do. But things seem different when it comes to the empirically tractable phenomena I have been discussing so far: even if we learn that neural activity reliably accompanies particular experiences, it *seems* reasonable to ask why such states feel the way that they do, or why they feel like anything at all. Many philosophers thus posit an epistemic or ontological gap between the mental states we experience, and the intentional and physical states that we can characterize empirically (Chalmers, 1996; Levine, 1983). An enormous literature has developed to address the nature of this gap, but Dennett calls this game off before it gets started. In light of his conventional realism, it should be unsurprising that he treats the experience of primitive, self-standing, first-personal states as an illusion that human brains reflexively create. But his revisionary proposal is not as radical as it first seems.

When I describe the subtle notes of vanilla and tobacco in coffee, I sincerely express my beliefs about how it tastes. According to Dennett, such self-reports should be treated as ordinary targets of empirical investigation. Sincere reports provide (fairly) unambiguous evidence about how things seem to a person. And we should generally believe someone who characterizes a current experience as seeming to have a particular character. We should also believe someone who claims not to be having a particular experience, even if it seems to us that they should be. And if we learn that someone believes that they are having experiences that it is impossible for them to have (e.g., if they are experiencing Anton's syndrome), we should look for an explanation of their beliefs, not an explanation of their nonexistent experiences. So far, so good. But Dennett also argues that our access to subjective experience is always conceptually mediated, and that we should acknowledge the possibility that there is nothing more to say about experience than that it is judged to be a particular way. This is where his position on consciousness begins to diverge from the view that is common among Anglo-European philosophers.

Dennett aims to replace the kind of lone-wolf auto-anthropology that treats claims about subjective experience as unassailable with a heterophenomenological method that treats judgments about subjective experience as (potentially false and potentially misleading) beliefs about how things seem to us (Dennett, 2003). While everyone acknowledges that I can misrepresent some characteristics of the coffee, Dennett holds that I might be wrong about how the coffee tastes to me. Many philosophers worry that focusing on beliefs constitutes a refusal to take consciousness seriously (cf., Chalmers, 1991). They claim that there is always a further question to ask about the qualitative states that lie behind our beliefs. And if we were to learn that qualitative states reliably cause our beliefs about subjective experience, this would force Dennett to accept a reductive and realist account of subjective experience. By rejecting this possibility, and claiming that we should focus on beliefs instead of experiences, Dennett appears to be adopting a behaviorist and verificationist perspective, and explaining consciousness away. But Dennett argues that he is taking consciousness as seriously as it can be taken. We can ask ourselves how we feel, and we will usually come up with a response. We can ask others how they feel, and they will tend to respond in kind. But how could we uncover qualia lurking behind these responses? Dennett claims that we could not: the only way we could know about qualia is through our beliefs about how things seem; and we have every reason to submit such beliefs to ordinary empirical investigation, and to recognize that our beliefs might be wrong here, just as they may be in any other case.

Of course, we can examine subpersonal states, using tools from the cognitive and computational sciences. But by definition, such states will not be

experienced by a person, and they will not be qualitative states; they will be states of information processing wetware, carrying out intentionally specified tasks. Even worse, subpersonal processing offers a picture that stands in stark contrast to everyday assumptions about subjective experience. Nothing comes prepackaged as a simple, fully determinate unity; mental states are constantly being constructed by dynamic neural assemblages, which pass in and out of existence as we move about in the world (cf. Barrett, 2017). Queries about subjective experience lead us to dress subpersonal states in person-level clothes, and the resulting forms of metacognition impose interpretive commitments onto our claims about how things seem to us. Thus we find it easy to smuggle ontological commitments into our judgments about subjective experience (Dennett, 2017). Simplifying somewhat, our reports lead us to posit primitive qualitative states because they privilege one stream of processing, and obscure the diversity and complexity of the subpersonal computations that generate moment-to-moment experience. In typically functioning brains, numerous processes operate in parallel to generate the stability that is required to guide moment-to-moment behavior, and to provide determinate shape for moment-to-moment experience. But without a Cartesian Theater where “everything comes together,” this unity of experience must be constructed by a diversity of subpersonal states and processes. And over the course of his career, Dennett has developed two useful metaphors for explaining how this occurs: “the multiple drafts model,” and “fame in the brain” (see Dennett & Akins, 2008, for a detailed introduction to these metaphors).

The multiple drafts model begins from the assumption that a typical human brain contains numerous systems that carry out tasks in parallel, in accordance with their evolved or acquired purposes. The information embodied in these processes rarely needs be bundled into a unified state to be used, because local discriminations are typically sufficient to guide fluid and adaptive behavior. But when we query the state of the brain—by soliciting a judgment or another behavioral measure—a univocal response is extracted. By binding the resulting representations together in working memory, we call subjective experience into existence. But there is “no reality of conscious experience independent of the effects of various vehicles of content on subsequent action” (Dennett 1991, p. 132). While we find it easy to assume that there is a fact about when contents “become conscious,” this is an artifact of the unity and determinacy that is imposed by linguistically structured judgments. And since there will always be ongoing streams of processing that we haven’t—and may never—formed beliefs about, questions like “when did I (as opposed to parts of my brain) become informed, aware, or conscious of some event?” may not have, or need, and answer; and they may even have false presuppositions (Dennett & Akins, 2008).

The hypothesis of “fame in the brain” elaborates upon this position, clarifying what it means for a cognitive state to “become conscious.” Fame emerges as an artifact of the competition for scarce attentional resources; so most people never become famous, and those who do are always at risk of falling out of favor when the next big thing comes along. But in every case, fame eventually fades, occasionally disappearing without a trace. Likewise fame in the brain is an artifact of computational competitions for scarce attentional and action-guiding resources; few cognitive states win these competitions, and those that do are always at risk of being overtaken by processes that are more salient in the current context. Representations that rise to the level of consciousness always fade away, and they often disappear without leaving any trace, unless there is some reason to record their fame for future use. Furthermore, just as people cannot be famous for only a minute (their fame must extend long enough to be publicly verified from multiple perspectives), conscious mental states must persist long enough to affect various streams of information processing, and they must be robust enough to recruit systems downstream to follow their lead (cf., Dehaene, 2014). Finally, just as there will rarely be a determinate moment at which someone becomes famous, there will rarely be a determinate point at which a neural state or process becomes conscious. We tend to see fame retrospectively, but once a person has become famous, all eyes are on them—and once a cognitive state has recruited attention and working memory, it can guide the unfolding of thought and action, at least so long as it retains control over the production and consumption of representational resources. This ongoing construction of fame in the brain generates the continuous “flow” of conscious experience.

The constructions that are generated as cognitive states “win” in the competitions for attention and the control of ongoing behavior are only available for particular uses, in particular contexts. These representations are important, as they allow us to “broadcast” facts about our current situation to others, and to revise our assumptions about what we should do next. But while the outputs of this process seem like self-standing entities, they are fragile, transient, unstable, and constantly undergoing revision. Conscious states pass out of existence when they are no longer relevant to ongoing behavior. This is the core of Dennett’s claim about subjective experience, and it is the respect in which he sees conscious states as illusory. Subjective experiences seem to be unified because the constructive and competitive processes that produce them are not available for further reflection; and any attempt at investigating how we feel will impose unity on multiple streams of computational processing so that we can report on our current state to ourselves and to one another. But there is nothing that answers to the apparent unity of subjective experience beyond our momentarily unified beliefs about the unity of our current subjective experience.



Against this backdrop, David Rosenthal's contribution to this volume explores the extent to which Dennett can privilege phenomena that are accessible from a third-personal perspective, explain our first-person access to mental states, and preserve the features of first-person experience that we have antecedent reason to accept. His primary target is Dennett's (1991) rejection of the claim that there is a "way things actually, objectively seem to you even if they don't seem to seem that way to you!" (p. 149). Of course, it's not always easy for us to figure out how things do seem to us. But Rosenthal argues that once we acknowledge that it can be difficult to figure out how things seem to us, we must allow for a distinction between seeming and seeming to seem. For unless seeming to seem is distinct from seeming, we won't be able to "distinguish conscious from subliminal perceiving, nor conscious from nonconscious believing" (p. 151). Rosenthal rightly notes that this would be a bad result. But he claims that a Higher Order Thought (HOT) approach to consciousness can make room for a distinction between seeming and seeming to seem, while preserving the most important aspects of Dennett's third-personal strategy for examining consciousness. Rosenthal's critical engagement with Dennett thus makes it clear that they are close allies, and it paves the way for more robust interactions between those who treat consciousness as an illusion and those who defend a HOT theory of consciousness.

In a very different way, Jesse Prinz also urges Dennett to retreat from some of his more contentious claims about the illusion of consciousness. However, he tries to cut out a middle path between Dennett's attempts at debunking inflationary claims about conscious experience and his own attempts to demystify consciousness while preserving it as a real, rather than illusory phenomenon. Prinz argues that visual imagery is unlikely to be encoded in either mental pictures or linguistic descriptions of such pictures (as Dennett has sometimes suggested that it should be); but like Dennett, he argues that it's unlikely that we will find a way to explain the phenomenology of visual imagery that will preserve the content and structure of the representations employed by subpersonal processes. Prinz then argues that conscious experience is richer than Dennett believes. On his view, consciousness arises much earlier in cognitive processing, prior to our judgments about what we are experiencing; but like Dennett, he acknowledges that we often think we are experiencing the world in richer detail than we actually are. Prinz then argues that when we start by asking whether conscious states share anything in common, we find that attention is at play wherever we find conscious experience. And he argues that his preferred account of consciousness—Attention to Intermediate-Level Representations—provides a way to naturalize qualia that even Dennett could love. This is an interesting proposal, though questions are likely to remain about whether Prinz has provided a way of explaining qualia, or has instead explained them away.



Finally, Andy Clark also takes up Dennett's claim that qualia are illusory. He does so from the perspective of the emerging consensus that brains use what they already know to predict future interoceptive and exteroceptive signals. In his recent work, Dennett (2014) has argued that we project our embodied expectations onto the world in ways that generate experiences of cuteness, sweetness, funniness, and other qualitative phenomena. Dennett claims that qualia are illusions that arise because we assume that these phenomena are features of the world that we are detecting; but we are just predicting our own embodied reactions to the world, and confirming these predictions, so there's no "real magic" to be found behind our judgments about how things seem. Clark largely agrees, but he worries that this framing of the hypothesis is unable to explain how we experience surprising qualia. According to Dennett, we experience qualia as real because we predict that we will encounter them. But we also experience unexpected things when we first encounter novel phenomena, and Dennett makes this seem more mysterious than it should be. Clark thus provides an amendment to Dennett's Bayesian hypothesis. He argues that the "raw sensory inputs actually possess plenty of hidden structure—so spotting a novel kind of cuteness is really no harder than spotting a novel kind of chair as a chair, or some crazy futuristic automobile as an automobile" (p. 214). Like Prinz, Clark thus suggests that qualia may be just as real as tables, chairs, and puppies—but maybe this brand of qualia realism shouldn't worry Dennett (especially if adopts the brand of deflationary realism that Kukla argues for in this volume).

#### 4. Darwinian Thinking

Turning from questions about consciousness, to questions about subpersonal content, Dennett (1975) argues that there are important structural similarities between the kind of intergenerational learning that occurs by way of natural selection, and the more familiar kind of learning that's governed by the behaviorist Law of Effect. He notes that heritable traits that help animals to survive and reproduce will typically accumulate within a population, while rival phenotypes will typically disappear. This yields a situation where animals adapt to the stable features of their local environment over many generations. Likewise, animals tend to repeat those behaviors that are followed by pleasant outcomes, and they tend to avoid behaviors that are followed by unpleasant outcomes; over time, this yields a form of behavioral learning that allows them to attune to salient patterns of rewards and punishments in their local environment. Both of these learning processes are self-correcting, and both draw on statistical regularities to shape future behavior. Just as importantly, both forms of learning yield regular forms of goal-directed behavior, which are readily interpretable from the perspective

of the intentional stance. Dennett's gradualist account of how complex forms of agency arise from simpler ones centers on the Tower of Generate and Test, which builds on this early insight.

At the bottom of the Tower of Generate and Test, we find simple organisms with evolved motivations that have been optimized to guide specific actions in specific environments. The successes and failures of these Darwinian organisms are preserved through differential reproduction, as successful organisms pass their dispositions along to the next generation, whereas unsuccessful animals do not. This yields a form of cross-generational learning, as ineffective motivations are eliminated, and selective pressures cause the heritable traits associated with reproductive fitness to accumulate within a population. Over time, Darwinian animals begin to act as if they were designed to maximize their chances of surviving and reproducing. And so long as their environments remain fairly stable, the research and development that's carried out by natural selection will often calcify in basic forms of biological preparedness (Cummins & Cummins, 1999). These animals do not think, they just act.

But the world we inhabit is dangerous and dynamic, and information about the best times and places to forage is often inaccessible to the mechanisms of natural selection. In this context, animals that make foraging decisions with higher payoffs tend to have an advantage over those that wander randomly. So many animals possess onboard behavioral guidance systems, which they can use to learn about local regularities, and to figure out where and when to eat, and where and when to mate. Many forms of behavioral guidance do not require subjective experience, even if they require decision making and even if they are goal directed. For example, the Skinnerian organisms that inhabit the second floor of the Tower of Generate and Test adjust their behavior in light of the rewards and punishments that accompany their actions, and they do so reflexively and automatically. Through repeated trials in a single life, such organisms can learn more successful ways of getting around in the world. And Dennett (1975) argues that this basic form of instrumental learning is the foundation for many forms of biological cognition, as surviving in a dangerous and dynamic world requires making predictions about the likelihood of rewards and punishments, and updating those predictions to better track the patterns of rewards and punishments in the world. This is significant because Skinnerian learning yields some degree of control over ongoing behavior, and it does so without requiring consciousness.

Instrumental learning is, however, limited by the statistical regularities that emerge in an agent's successes and failures, and by the forms of biological preparedness that an organism inherits from its ancestors. While Skinnerian organisms can look to their past experiences to decide which action to take next, this decision-making strategy comes up short where mistakes are costly, or even

worse, deadly. Selective pressures have thus pushed some animals to learn to think ahead. As Dennett (1996, p. 133) argues, an animal that can do this can learn more rapidly and more accurately, all while decreasing the likelihood of costly (even deadly) mistakes. By building models of the world they inhabit, and using these models to think about what they should do next, the Popperian animals on the third floor of the Tower of Generate and Test can detect some mistakes before they make them; and just as importantly, they can compare the potential rewards of carrying out different actions. The ability to let your hypotheses die in your stead would yield a massive selective advantage for any animal that happened to hit upon this neat trick; so Dennett argues that we should expect many animals to possess at least rudimentary Popperian capacities. And there is an emerging consensus that many animals have a great deal more behavioral flexibility. Crypto-Popperian agents may be found deep in the phylogenetic tree; and recent examinations of expectation-driven learning and prospection suggest that even mice may build forward-looking models of the world (Seligman, Railton, Baumeister, & Sripada, 2013). This is not an issue that I can explore here, but Dennett (2017) addresses it in detail, and argues that understanding the pervasiveness of expectation-driven learning is likely to transform our thinking about mentality! But no matter how common this kind of processing is, there is still likely to be massive variability in the capabilities that are possessed by different Popperian agents.

The most sophisticated Popperian organisms take this capacity to plan ahead to its logical conclusion: they think about what they should think about next. According to Dennett, creatures that can do this are the only ones that we can be sure have subjective experiences, as their conscious representations play a critical role in the ways that they plan alone, and in the ways that they plan together. Using a variety of mind-tools—including language and other forms of social exchange—these Gregorian animals are able to act in ways that allow them to improve the internal environments they use to plan ahead. Their method of reasoning remains fully Popperian, and this preserves the continuity with other forms of mentality. But these organisms are able to compare merely possible notional worlds, and not just expected material outcomes. Put differently, they can compare their ideas about how things seem against their ideas about how to make things different. Their plans can open up genuinely new possibilities. And by representing their reasons *as reasons*, they can engage in directed forms of niche construction.

In his contribution to this volume, Peter Godfrey-Smith examines Dennett's claim that there's a relatively linear progression from Darwinian organisms to Gregorian organisms. He argues that Darwinians shouldn't expect to find a linear

ordering like Dennett's Tower of Generate and Test; they should expect to find a tree with multiple branches that share some features and differ in many others. Of course, he acknowledges that many trees have shared central structures. So he begins by examining a class of associative learners—the Humean organisms—that lie between Darwinian and Skinnerian organisms. Humean organisms can learn through classical conditioning, but they don't display capacities for self-directed learning. Still, the capacity to learn by way of instrumental conditioning appears to have been built upon the capacity to learn by way of classical conditioning. And this suggests the beginnings of a tower: all Humean organisms are Darwinian organisms, but not vice versa; and all Skinnerian organisms are Humean organisms, but not vice versa. But things rapidly become more complex, as there are many potential routes from Skinnerian to Popperian cognition, and there are many cognitive capacities that are more demanding than instrumental conditioning, but are not fully Popperian. For example, Godfrey-Smith posits Carnapian organisms that can extract patterns and extrapolate upon them—this allows them to partially decouple their representations from stimuli, even if they cannot plan ahead. He also posits a class of Pearlman organisms, which can engage in forms of causal reasoning, even if they do not plan ahead, and even if they do not extract patterns and extrapolate from them more generally. Minds can also be social, without being Popperian; and Godfrey-Smith argues that multiple mimetic organisms have found ways to capitalize on the R & D carried out by others, by copying successful actions and avoiding unsuccessful actions, all without developing the capacity to plan ahead. Finally, there appear to be Tolmanian organisms, which rely on cognitive maps, without using them to run hypotheses offline.

While Godfrey-Smith's discussion of these different kinds of minds is incredibly helpful, the most important aspect of his chapter is his claim that different kinds of minds can be nested, stacked, and combined with one another. This provides a way of understanding the cognitive diversity across the phylogenetic tree; and the possibility space he has articulated is a playground for philosophers. At the beginning and at the end, this project remains largely consistent with Dennett's picture. But seeing cognitive capacities as embedded within one another, or operating alongside one another, has important implications for thinking about neurodiversity, and for thinking about the kind of Darwinian compatibilism that Dennett defends. The top of the tree, however, is where Darwinian humanism hits its stride. Dennett argues that Gregorian animals possess a distinctive kind of freedom, which is worth wanting; and he argues that they do so because of evolved capacities for thought, action, and consciousness.

## 5. Darwinian Compatibilism

A laboratory-bound artificial intelligence can be pre-programmed with any information it will need to keep its batteries charged. But biological systems are rarely so fortunate. Many of them possess dispositions like those that helped their ancestors to survive and reproduce. And the extraordinarily lucky ones possess abilities to build upon those capacities to improve their lots in life. As I suggested in section 4, Dennett argues that natural selection has opened up varying degrees of freedom for different sorts of organisms. Many philosophers have started to take note of the data from the cognitive and biological sciences suggesting that the human brain is a predictive behavioral guidance system, which mines past experience for information that allows it to improve its future predictions (for very different overviews, see Clark this volume, and Huebner this volume). But from his earliest critiques of Skinner, to his most recent thoughts about practical agency, Dennett has been arguing that our ability to form forward-looking expectations has a significant impact on the kind of freedom we possess.

Since the early 1970s, Dennett has been developing an account of human agency that explains how human freedom arises in biomechanical systems, and persists in a world governed by natural selection. We are resource-limited agents who must navigate dynamic decision-making environments in a timely manner; and we must often plan for the future because the social and material environments that we inhabit make it impossible to micromanage every detail of real-time deliberation. This is why “the policy of preparing oneself for tough choices by arranging to be determined to do the right thing when the time comes is one of the hallmarks of mature responsibility” (Dennett, 2003, p. 117). By planning ahead, we can organize the details of our local environment to make decision-making fast, effective, and reliable. But the core of Dennett’s Darwinian humanism arises from his claim that humans *collectively* possess the skills and capacities that are necessary to open up greater degrees of freedom; we do so through a process of cultural niche construction, which allows us to capitalize on intergenerational and individual forms of learning.

An initially seductive, though not implausible hypothesis about the mechanisms of cultural change, relies on Great Minds, who are so intelligent that they can see possibilities that no one else can see. But for a Darwinian, intelligence is always situated, embodied, and interpretable only relative to the options that the world affords for a particular entity. A smart cat knows how to acquire food when it wants to; a smart dog, or a smart crow, may know how to capitalize on the affections of nearby humans; and a smart person will generally know how to navigate the socially structured world. We do a lot of work managing what we will treat as intelligence, and we privilege certain kinds of trajectories through

social space. But when someone succeeds at a locally popular game, this often tells us little more than that the person has been trained to play that locally popular game well. No doubt, a high degree of competence is required to play any game well; but if we are looking for innovation, and the opening up of novel possibilities, then we are going to have to look elsewhere. As Dennett notes, we should look for cultural novelty in much the same place that we look for biological novelty. It emerges through countless trials and errors, through the weeding out of bad ideas, and through exploratory attempts spread across multiple people doing multiple different things. And at the end of the day, cultural evolution is often just as mindless as biological evolution—but both can produce distinctive forms of freedom, or so Dennett argues.

In his most recent book, Dennett (2017) develops a comprehensive account of how the memetic transmission of ideas makes human culture and human freedom possible. He argues that our social environments afford opportunities for action, which open up the possibility of social forms of freedom. Most nonhuman animals inhabit evaluative landscapes that are structured *exclusively* around primary rewards like food, water, and mating opportunities; and many animals are able to make decisions that will allow them to increase their chances of success in pursuing rewards and avoiding punishments. These animals possess a form of freedom: they can decide which trajectory to cross in the pursuit of the rewards that they happen to care about. But such animals are not free to decide which ends to pursue, and they are not free to decide that it would be better to give up the goals that they have inherited. Most humans possess these forms of freedom as well, but even here, our evaluative landscapes tend to be richer and more varied. We possess evolved preferences for conformity and prestige, and these allow us to treat almost any perceptible phenomena as valuable. We praise and sanction one another, and internalized evaluative expectations provide us with a set of open options over which we can deliberate, giving us the freedom to decide which of our available ends we should pursue, and which we should avoid.

Still, this might not seem like much freedom. But the options open up quickly. As we move through the world, we acquire habits of thought, and habits of behavior that can guide us through familiar landscapes. But Gregorian creatures can also construct models of how the world could be, and use these to models to consider the feasibility and desirability of pursuing novel possibilities. In an important sense, this is a simple extension of the kind of freedom we find in non-human animals: we can choose between the options that are available in our evaluative landscape; but our evaluative landscape is filled with some *mere* possibilities, and we can choose to pursue those possibilities even though they only exist notionally. Having chosen to examine a notional possibility, we can then attempt to bring the world we represent into existence. Importantly, every action

structures the environment we inhabit, making certain options more transparent, and others harder to perceive. For example, chefs quickly learn that the physical organization of a kitchen can make it easier to cook some things, and more difficult to cook others; our everyday use of money makes some forms of exchange much easier, but in doing so it makes some forms of mutual aid and mutual support more difficult to sustain; and the patterns of gentrification and white flight in a city make certain forms of racialized thinking more prevalent, as they affect the way that we conceptualize the space of socially available actions. But just as we construct social niches to make our task easier, we can use resources that emerge in these niches to open up novel possibilities. Of course, we rarely do this on our own. But linguistic representations allow us to broadcast our ideas to others; and we can build material structures that express our ideas in a form that is publicly consumable. And once we have made our ideas concrete, they become part of the world we can manipulate, and part of the world that others experience.

Of course, very few ideas are good ideas. Many are ignored, many are quickly forgotten, and many fall out of favor when the next big thing comes along. But where materialized ideas stabilize within a population, they can begin to open up new possibilities, and new opportunities for action. The fact that thought becomes crystallized materially and socially provides the key to Dennett's Darwinian theory of cultural evolution; he claims that learning how to make *better* plans requires finding a way to capitalize on the socially structured facts about our environment that we track as reasons. In their contribution to this volume, Luc Faucher and Pierre Poirier develop an account of memetics that recognizes a wider variety of evolutionary algorithms, and yields a more pluralistic approach to cultural change. They consider the adaptive characteristics of the human immune system, and they argue that the evolutionary algorithms that are at play here are not strictly matters of replication. They then turn to recent work in genomics, and they contend that would be a mistake to think that there is a single source of evolutionary information (encoded in the genome; or encoded in the memome) that carries the same content in every environment. Building on these discussions of biological phenomena, they defend an approach to memetics that relies on the operation of multiple selective "algorithms, operating at different time-scales on different units of selection and with different logical structures" (p. 267). In the final part of the paper, they draw together resources from David Amodio's account of stereotype representations, and Lawrence Barsalou's situated and embodied theory of concepts to show how different kinds of selective processes can be integrated to yield stable cultural phenomena.

Unlike nonhuman animals, many humans inhabit numerous partially overlapping, but subtly different micro-worlds. These microworlds have their own patterns of rewards and punishments, and they require using subtly different forms



of intelligence to navigate them successfully. So people will often form multiple situation-specific strategies for coping with environmental contingencies. At the extreme, hierarchical political institutions can lead people to rely on DuBoisian forms of double consciousness to navigate a fractured world; but the environments we inhabit are never hegemonic, and there is always room to decide which kinds of relations to pursue, and which ends to take up and abandon—though each choice has its costs. Put differently, most of us are free to choose which of several ends we would like to pursue. The possibilities that we perceive as open and available then become the raw resources that we use in constructing the narratives that we tell about ourselves, and they provide the foundation for thinking about what is possible for us. And in some cases, we may even be able to combine resources from different microworlds to open up further forms of freedom. To the Darwinian humanist, this feels like all the freedom that's worth wanting.

These phenomena provide the foundation for my own examination of Dennett's account of human freedom. In my contribution to this volume, I outline an account of how humans can collectively construct novel possibilities, and argue that discussions of agency and freedom should focus on issues of micropolitics, instead of issues in metaphysics. I proceed by considering points of convergence between Dennett's account of human freedom, and Michael Bratman's account of planning agency, and its role in intertemporal and interpersonal coordination. First, I examine the claim that foreknowledge can give us control over ongoing behavior; by reverse engineering the capacities that would be required for planning agency, I then argue that individual freedom is typically constrained by the statistical structure of a person's social and memetic environment. And I provide support for Dennett's claim that since freedom is a person-level phenomenon, it is often sustained by help from our friends. Using Bratman's approach to interpersonal coordination, I develop an account of human freedom that is socially scaffolded, and that requires listening and learning from others. While this chapter only scratches the surface, it provides a framework for thinking about freedom and agency as interpersonal phenomena.

## 6. What if Dennett Is Wrong?

This brings us to the themes that structure the final contribution to this volume. Lynne Rudder Baker develops several lines of criticisms against Dennett's attempt to break the "spell" that prevents people from submitting their religious beliefs and practices to scientific investigation. She argues that religion, as such, is unlikely to be a unified phenomenon, and she contends that the drive to treat religion as a unified phenomenon leads Dennett to defend an implausible memetic approach to religion, and to mistakenly suggest that the existence of a Hyperactive Agency



Detection Device would impugn religious belief. While Baker agrees with that Dennett that religious beliefs and practices should be studied scientifically, she argues that it would be a mistake “to suppose that science is the exclusive arbiter of genuine reality (i.e., of what entities, properties, and kinds exist irreducibly and ineliminably)” (p. 342). For Baker, accepting Darwin’s strange inversion of reasoning leaves us without a story about why people deserve to be treated with dignity because they are people. And she claims that Dennett’s commitment to an anti-essentialist approach to persons makes human dignity a matter of convention (and this is bad news, given the susceptibility of the intentional stance to group-based considerations and given the biased practices that guide our everyday patterns of reasoning). If we want to salvage human dignity, Baker argues, we must see people as possessing a first-person perspective *essentially*; and she claims that any theory that suggests otherwise is likely to be driven by an unsubstantiated ideology.

Of course, Dennett is unfazed by these criticisms. He’s a committed Darwinian who thinks humans are exceptional precisely because of their evolutionary history. He also thinks that appeals to essences and dignity are ungrounded, unless they are consistent with a gradualist theory of biological and cultural evolution. This much will be familiar to anyone who knows Dennett’s work. His replies to these critical arguments are unlikely to convince Baker; but this is where the defense of Darwinian humanism must take place. I could say more about this, but I should probably stop here, and put those ideas in my own book manuscript.

## Works Cited

- Akins, K. (1996). Of sensory systems and the “aboutness” of mental states. *Journal of Philosophy*, 93, 337–372.
- Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Boston, MA: Houghton Mifflin Harcourt.
- Chalmers, D. (1996). *The conscious mind*. Oxford, UK: Oxford University Press.
- Cummins, D. D., & Cummins, R. (1999). Biological preparedness and evolutionary explanation. *Cognition*, 73(3), B37–B53.
- Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. New York, NY: Penguin.
- Dennett, D. C. (1975). Why the law of effect will not go away. *Journal for the Theory of Social Behaviour*, 5(2), 169–188.
- Dennett, D. C. (1987). True believers. In Daniel C. Dennett, *The intentional stance* (pp. 13–42). Cambridge, MA: MIT Press.
- Dennett, D. C. (1991). *Consciousness explained*. New York, NY: Little, Brown, and Company.

- Dennett, D. C. (1996). *Darwin's dangerous idea: Evolution and the meanings of life*. New York, NY: Simon and Schuster.
- Dennett Daniel, C. (2003). *Freedom evolves*. New York, NY: Viking.
- Dennett, D. C. (2007) Philosophy as naïve anthropology: Comment on Bennett and Hacker. In M. Bennett, D. C. Dennett, P. M. S. Hacker, & J. Searle (Eds.), *Neuroscience and philosophy: Brain, mind, and language*. New York, NY: Columbia University Press.
- Dennett, D. C. (2014). Why and how does consciousness seem the way it seems? In *Open MIND*. Frankfurt am Main: MIND Group.
- Dennett, D. C. (2017). *From bacteria to Bach and back*. New York, NY: W. W. Norton and Company.
- Dennett, D. C., & Akins, K. (2008). Multiple drafts model. *Scholarpedia*, 3(4), 4321.
- Drayson, Z. (2012). The uses and abuses of the personal/subpersonal distinction. *Philosophical Perspectives*, 26(1), 1–18.
- Drayson, Z. (2014). The personal/subpersonal distinction. *Philosophy Compass*, 9. DOI: 10.1111/phc3.12124.
- Friston, K., & Frith, C. (2015). Active inference, communication and hermeneutics. *Cortex*, 68, 129–143.
- Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naive theory of rational action. *Trends in Cognitive Sciences*, 7, 287.
- Huang, J., & Bargh, J. (2014). The selfish goal: Autonomously operating motivational structures as the proximate cause of human judgment and behavior. *Behavioral and Brain Sciences*, 37, 121–135.
- Huebner, B., & Rupert, R. D. (2014). Massively representational minds are not always driven by goals, conscious or otherwise. *Behavioral and Brain Sciences*, 37, 145–146.
- Huebner, B. & Yergeau, M. (2017). Minding theory of mind. *Journal of Social Philosophy*, 48(3), 273–296.
- Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, 64, 354–361.
- Lycan, W. (1990). The continuity of levels of nature. In W. S. Lycan (Ed.), *Mind and cognition: A reader* (pp. 77–96). Oxford, UK: Blackwell.
- McGeer, V. (2007). The regulative dimension of folk psychology. In *Folk psychology reassessed* (pp. 137–156). Dordrecht, The Netherlands: Springer.
- Roepstorff, A., & Frith, C. (2004). What's at the top in the top-down control of action? Script-sharing and “top-top” control of action in cognitive experiments. *Psychological Research*, 68(2–3), 189–198.
- Rosenberg, J. (1994). Comments on Bechtel, “Levels of description and explanation in cognitive science.” *Minds and Machines*, 4(1), 27–37.
- Seligman, M. E., Railton, P., Baumeister, R. F., & Sripada, C. (2013). Navigating into the future or driven by the past. *Perspectives on Psychological Science*, 8(2), 119–141.

- Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, 18, 186–193.
- Spaulding, S. (2017). How we think and act together. *Philosophical psychology*, 30(3), 298–314.
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59, 433–460.
- Wegner, D. M., & Gray, K. (2016). *The mind club: Who thinks, what feels, and why it matters*. New York, NY: Viking.
- Zawidzki, T. (2013). *Mindshaping: A new framework for understanding human social cognition*. Cambridge, MA: MIT Press.



## **PERSON-LEVEL AND SUBPERSONAL EXPLANATIONS**



# 1.1

## EMBODIED STANCES

### REALISM WITHOUT LITERALISM

Rebecca Kukla

Daniel Dennett's work on the intentional stance had a seismic effect on philosophy of mind in the late 20th century. Navigating between the unabashed inflationary realism of the defenders of a language of thought and the spare ontology of the eliminativists, Dennett's suggestion was that beliefs, desires, and other such attitudes were the ontological complements of a certain sort of *strategy*. A system has beliefs and desires just in case taking the intentional stance toward that system—explaining and predicting its behavior—is strategically effective. The intentional stance is Dennett's biggest stance hit. But he also appeals to other stances, such as the *design stance* (in which we take a system as functionally or teleologically defined, as we typically do with machines, plants, and body parts) and the *physical stance* (in which we take a system as merely causal).

An evident question to ask about Dennettian stances is, what is the ontological status of the kinds of entities they appeal to? If one is a stance fan, does one believe that beliefs and desires are objectively real, or merely instrumental, or some other variety of real? Dennett has taken up this question many times, but, as we will see, his answers are often hard to interpret and sometimes inconsistent with one another. Another fundamental question about stances that has received less attention is, what is it, exactly, to take up or occupy a stance? It is somewhat surprising how little this question has been discussed, given that it is clear that a stance is a kind of *pragmatic* posture of some sort, governed by certain kinds of purposes. As such, we need to ask what sort of pragmatic activity it is. Is it merely the intellectual entertainment of a set of questions and explanations, for instance? I will be arguing that taking up this later question, and seriously thinking about what we are

*doing* when we adopt a stance, is crucial to answering the first question, about the ontology of the objects of stances, in a satisfactory way.

I will argue that we ought to take stances not as merely intellectual attitudes, but rather as collections of concrete strategies for coping with objects and coordinating with others. These strategies will be embodied; we should take seriously the idea that a stance is, first and foremost, a way of readying your body for action and worldly engagement. The entities that show up from within a given stance are loci of norm-governed behavior, resistance, and explanatory power. The *real* things, from any stance, are the things we grapple with.

I claim that there is no *separate* question to be asked coherently as to whether these “real” entities are *really real* or *literally real*. The point is twofold. First, the idea that we can ask and answer this presumes some sort of supstance—a stance outside all stances—from which the entities that we cope with from a given stance can be compared to some sort of stance-independent objective reality. Once we understand what stances are, I argue, this presupposition ceases to make sense. Second, I argue that the notion of the literally real itself only gets a grip from within a specific stance—one that is *not* the intentional stance, but is rather what I will dub the *interpretive stance*. Outside of the interpretive stance, I argue, questions about the literal reality of intrastance entities generally deflate to practical questions about the success of various coping strategies.

## 1. Dennett on Stances and Their Ontologies

As is well-known, Dennett’s most distinctive contribution to the philosophy of mind is his argument that propositional attitudes such as beliefs and desires *just are* the states of systems that allow us to explain and predict those systems by taking up the intentional stance. He defines the intentional stance as a set of strategies; for instance, he writes, “First you decide to treat the object whose behavior is predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this agent will act to further its goals in the light of its beliefs” (1987, p. 17); the “fundamental rule” of the intentional stance is that the one taking it up must “attribute the beliefs [and desires] the system *ought to have*” (p. 20). But when is taking this stance *correct*—that is, when is the stance-taker *right* to attribute beliefs and desires to a system? Famously, Dennett’s answer is that she is right to do so *exactly when adopting this strategy yields explanatory and predictive success*. Thus the ontology of beliefs and desires is inextricable from the pragmatic success of attempts to attribute them. While the intentional stance is most useful for elaborate human-like systems, Dennett also thinks that it works just fine

on most animals, including simple ones like clams, and even on many artifacts and plants. By his reckoning, these systems *really do* have propositional attitudes, although this isn't often a very useful fact about them, and their system of beliefs and desires is impoverished.

This answer immediately raises large ontological questions. Is Dennett claiming that beliefs and desires *really exist*, even though the criterion of their existence is dependent on the pragmatics of attribution? Is he claiming that they are mere instrumental or theoretical entities that we appeal to in our explanations but that have no literal existence? Or is he carving out some third option? Evidence for each can be found in his writings over the years. Often, Dennett seems to want to be a realist (of some variety) about propositional attitudes while at the same time saying that their reality is inextricably bound up with or somehow constituted by our strategic ends. For instance, in "True Believers" (chapter 2 of *The Intentional Stance*) he writes:

My thesis will be that while belief is a perfectly objective phenomenon (that apparently makes me a realist), it can be discerned only from the point of view of one who adopts a certain *predictive strategy*, and its existence can be confirmed only by an assessment of that strategy (that apparently makes me an interpretationist). (1987, p. 15)

Here it certainly sounds as if the stance is what gives us *epistemological* access to objectively real phenomena that do not depend directly for their reality on the stance that reveals them. But a few pages later he says more cryptically, "All there is to being a true believer is being a system whose behavior is reliably predictable via the intentional strategy" (p. 29). This is quite a strong claim; it does not merely say that being a true believer (or really having beliefs) is *coextensive* with being reliably predictable from the intentional stance, but indeed that this is *all there is* to the reality of beliefs. What he seems to be equivocating on here is whether our explanatory stances are roughly successful *because* the entities they range over are real, or whether they are real *because* we take up and employ the intentional stance. The latter seems suspiciously idealist, and the former raises the question of what kind of reality they have that makes our stance-taking metaphysically appropriate.

In "Real Patterns" and elsewhere, Dennett (1991) argues that it is the *patterns* of folk psychology that are objectively real, despite their multiple realizability in different physical systems. He uses analogies such as centers of gravity, voices, and his infamous "lost sock center"—the point at the center of the smallest sphere containing all his lost socks—to show that all sorts of things that exist only as patterns still strike us as having unproblematic reality, even though they are not



*substantial objects* or parts of the fundamental furniture of the universe. Of where this leaves him on the question of realism, he writes:

My *ism* is whatever *ism* serious realists adopt with regard to centers of gravity and the like, since I think beliefs (and some other mental items drawn from folk psychology are *like that*—in being *abstracta* rather than part of the ‘furniture of the physical world’ and in being attributed in statements that are *true* only if we exempt them from *a certain familiar standard of literality* [my emphasis])” (1991, p. 72).

In the same chapter, he dubs this position “mild realism.” However, he tells us neither how to identify the fundamental furniture of the world as opposed to less substantial patterns, nor why we should accept that the physical furniture is the most fundamental. Absent some clear measure of fundamentality, it is not clear to me how it is helpful to grade realisms in terms of their *strengths*; I am not sure what is clarified by calling this realism “mild,” or what it means to call something real but not fundamental, or not a “thing” as he puts it later in (p. 114). It’s hard not to feel that Dennett is just using language here that he hopes will get him off the hook for committing to either straight-ahead realism or straight-ahead instrumentalism.

Also of crucial note for me here is Dennett’s explicit denial, in the quotation just given, that statements about beliefs and desires are *literally true*, or at least *literally true* by “a familiar standard,” whatever that means. This denial seems to me to fit uneasily with his claim in “True Believers” that “belief is a perfectly objective phenomenon.” I am not sure what sorts of things are supposed to be real or objective but not a topic for literally true claims, nor what it means when he says that instead of being literally true, such claims can be “true with a grain of salt” (1987, p. 72). This literality standard will be crucial for me later.

We see a similar equivocation or perhaps a failure of commitment in his much more recent paper, “Kinds of Things: Towards a Bestiary of the Manifest Image” (2013). Here he points out a wide range of entities that are perfectly good candidates for thinghood in our “casual ontology,” including miles, sakes, cahoots, holes, and dollars, among others. At points, he apparently insists that the only real metaphysical questions we can ask do not concern whether such things belong to “a univocal, all-in, metaphysical truth about what there is” but rather whether there are parts of strategies that “help people move between different ontological frameworks,” and form robust parts of our everyday holistic metaphysical network (pp. 105–106). And yet, he worries that such entities “sit awkwardly in the world of atoms and molecules” and that studying them

allows us to evade rather than answer “the ultimate ontological question” by letting us stick to what is “believed-true” rather than actually true (p. 106). Such comments suggest that the study of such “casual ontology” reveals something *other* than the literally or fundamentally real entities. When he proposes treating “metaphysics as folklore studies” (p. 101), then, it is not clear whether he is getting rid of the distinction between strategic ontology and the “really real” or just suggesting that attention to the former is more interesting and productive for philosophers.

This is where Dennett leaves us. My strategy will be to come back to the ontology of propositional attitudes in an indirect way; I think that we need to say more about what is actually involved in *adopting a stance* or *employing a strategy* that is characteristic of a stance, and that doing so will allow an ontological story to fall out.

## 2. Performing Stances: Coping and Coordinating

For over a decade, I was married to a fellow philosopher. Typically, since we are both human beings, he took the intentional stance toward me: he treated me as an agent with a mental life whose propositional attitudes were more or less rationally connected to one another. However, I do not do well when I am hungry. I need to eat every couple of hours at the very most, and if I don’t, my rational capacities and my physical and emotional resilience degrade quickly. This happens so fast that I am often not aware of it. When I would start to functionally degrade due to hunger, my philosopher-husband would shout, “Design stance!” He would immediately stop trying to have rational conversations with me and basically drop everything to find me food, sometimes above my protests. As he would explain, at those times it was much more productive to treat me as a machine in need of repair and maintenance than as a rational negotiator of the space of reasons and social interactions. My *physical* structure did not dramatically change as I became hungry and then again as I ate. Yet he was clearly right that different strategies involving some pretty basically different presuppositions about what sort of system I was—what sort of pattern I incarnated—were appropriate at different times.

Although this husband was a philosopher and so explicitly named his stance-shifts, mostly these shifts were not intellectual exercises in taking up different explanatory or predictive attitudes. In the first instance, what changed were his *embodied coping strategies* when it came to coordinating with me. He changed his actual immediate plans, his tone of voice, his conversational goals, and much more, accordingly. There is clearly a thick-bandwidth, interdependent relationship between his coping strategies and which information about me could show

up as salient. This is so even though his different strategies did not magically alter my structure, and even though his strategies were fallible—on any given occasion he could be wrong about what the most effective strategy would be. (There are few things as infuriating as having your partner switch over to the design stance when you feel like your rational capacities are fully engaged!)

More generally, outside of the practice of philosophy, most of us don't spend much time self-consciously employing the intentional stance. Most of our belief and desire attributions are built into our embodied responses to the people around us. Our adoption of the intentional stance makes us more likely to ask someone to move rather than to pick them up and place them to the side, to inform the person running top-speed toward the bus stop that we know there is another bus on its way shortly, and so forth. None of this requires that we conceive of ourselves as engaged in folk-psychological explanatory projects.

My suggestion, in short, is that we need to take the word "stance" much more seriously than Dennett himself does, and to understand stances as, in the first instance, systematic collections of *embodied, performed strategies for coping and coordinating* with the world and the people in it. Outside of philosophy, a *stance* is a kind of a physical posture that readies our body for some sorts of reactions, activities, and perceptions, while making others difficult or impossible. Our bodily stance will shape what features of the world can be salient to us and how we can and cannot engage with it.

Consider, for instance, what is involved in taking up a boxing stance. Doing so gives one a specific kind of stability and range of motion; it protects certain parts of the body and leaves others free to move. The stance allows one to engage with another body, with the ring itself, with the bell, and so on, in specific ways. Once one has practiced boxing and developed some facility at it, taking up this stance also creates a set of implicit and explicit expectations for how things around will act and react. A proper boxing stance is embodied, but it also goes along with a whole set of intellectual configurations of attention, expectation, and strategizing. Adopting this stance may or may not be a successful strategy for coping with one's environment. It will work pretty well if an opponent begins boxing, but will be a miserable failure if an opponent instead pulls a gun or starts tap dancing.

Dennett's stances are strategies, and once we focus on how they are performed at the level of the body, it becomes clear that these strategies are not just lists of rules or heuristics sitting statically in the head. The intentional stance, or any other stance, I claim, begins in such an embodied posture, albeit one much less ritualized and narrow than the boxing stance. In fact, we can now see that the term "stance" suggests something overly static and synchronic. A bodily stance is just a starting position, not really an entire "stance," albeit one that interestingly constrains the worldly engagements that follow it. On my reading, "stances"

are systematic collections of embodied strategies for coping and coordinating; in fact, any Dennettian “stance” will in fact be made up of a wide and counterfactually flexible repertoire of bodily positions.

When we adopt an intentional stance and take someone as an intentional system, we ready ourselves to perceive certain kinds of data. We hear the sounds they make as speech and are attuned to their communicative efforts. We focus our eyes on their face and hands. We move our bodies differently around beings we take to have beliefs and desires. Our motions build in an expectation that theirs will coordinate with ours in predictable ways (not simply smashing into us, making eye contact, moving away from unpleasant or dangerous things, etc.); we use gestures to ostend things to them, and facial expressions to communicate with them. We reflexively take their actions as purposive and respond accordingly. When my ex-husband shifted from one stance to the other, his bodily posture changed—he shifted his attunement from the content of my words to their pitch and tone; he became actively directed toward the goal of obtaining food rather than toward engaging me in philosophical conversation, and so forth.

Compare what information is available to me if I take the intentional stance toward you, with what information would be available if I took what we might call the *clinical stance*. If I am a doctor examining your body in a clinic, I take it as a potential site of injury, disease, and various kinds of risks, as well as an entity to be assessed by its heartbeat, BMI, glucose levels, and so forth. This goes hand in hand with various embodied strategies; I touch your body in certain ways, perceive it from specific angles and through various machines. I encounter it in a specific sort of room and posture. Assuming you are conscious, there will be some mixing of the intentional and clinical stance; for instance, I need to be attuned to your expressions of emotion and physical discomfort, and I communicate with you about your symptoms and the like; but large swaths of your beliefs, desires, social relationships, and so forth, are hidden from view by the setting and the forms of interaction, and indeed, they ought to remain that way as they are simply not on the table, as it were.

We can see that three types of things are essentially complementary: (a) stances, or sets of coping strategies and expectations that are first and foremost embodied postures and performances flowing from these postures; (b) kinds of information made available by these stances, and (c) kinds of entities that are the bearers of this information and are the objective targets of the coping strategies that make up the stances. So the clinical stance is directed toward entities such as diseases, bodies, and wounds, and it yields information about viral loads and displaced joints. The intentional stance is directed toward entities such as rational agents and their psychological states and yields information about beliefs, desires, and other propositional attitudes.

I don't see any reason to think that Dennett would disagree with anything I have said so far. It is far from his style to overly intellectualize mental processes, or to assume that they take the form of explicit routines in the head. I have high hopes that he would agree that stances—which are, after all, *pragmatic* strategies by his own account—are in the first instance embodied, performed sets of expectations, coping tactics, and so forth. My goal so far is to draw attention to and emphasize this embodied dimension of stances, so that we can examine its significance, and see if it helps with the previous ontological questions.<sup>1</sup>

This emphasis, I think, helps us give more bite to Dennett's responses to questions about the ontological significance of the things toward which we are directed in stances. For instance, in an interview on naturalism, he says,

We have got all these atoms, and then we have the patterns that we discern among these atoms and four dimensions: space and time. Now the question is: Do the patterns have ontological significance? And for me the answer is: That's what ontology is. What other criterion could you ever use? What other reason could you ever have for your ontological presuppositions?" (Mirolli, 2002, p. 5)

This response can invite a reading in which "discernment" is a purely intellectual activity—an inner act of attending. Read in this way, his answer seems to collapse ontology into epistemology. As I read stances, however, the relevant kind of discernment—of patterns of rational connection between beliefs and desires, for instance—is a concrete part of coping. We discern what we need to respond to and what we need to navigate. As such, the concrete character of things richly constrains how we cope with them; our discernment is intimately enmeshed with the world itself. Not every pattern that is in some sense intellectually discernable is equally

---

1. Our readiness to take in and evaluate certain kind of information and to react in certain ways is sometimes embodied most saliently at the level of the brain and its chemistry. For instance, to borrow an example from Bryce Huebner (2015), some kinds of positive interactions with others that "induce mirth" trigger a release of dopamine that in turn changes how we search for and evaluate new information: "Positive affect induction often leads people to consider a broader range of contextual factors, and to be more flexible in trying different strategies for solving problems that antecedently interest them; it can also moderate anchoring effects, and increase the willingness to revise judgments" (p. 428). Such conditions at the level of the brain are certainly still embodied, and I don't see why we should not count them as part of the *posture* of the body. The conditions will only have the effects Huebner discusses as the body interacts dynamically with its ecological environment; in other words, if we were not *also* engaged in bodily macro-performances, we would not be able to identify systematic effects of the sort he describes. They are thus components of stances, preparing the body to engage with the world around it in specific ways.

real, because only some patterns really constrain our engagements. If we want to fix a bicycle, we have to take the design stance toward it, and from this stance its different functional parts will show up as salient and drive the relevant cuts in our behavior toward it. If we want to make it through a busy grocery store line, we generally have to take our fellow embodied humans as having beliefs, desires, frustrations, and hopes in order to navigate the line without things descending into a brawl. From within a stance, the *real* things, the things with ontological significance or heft, are the things that help constrain our negotiation of them in counterfactually stable ways, including the ability to thwart our plans and expectations when we misjudge them. To be real is just to be patterned, counterfactually robust in the sense that its behavior is predictable under a range of conditions, and resistant. Furthermore, nothing can be counterfactually robust or resistant except relative to some set of performed strategies for coping and coordinating with it.

Consider some of Dennett's (2013) examples of little-discussed ontic types in "Kinds of Things." There he insists that holes, voices and money, for instance have a more robust claim on reality than sakes, cahoots, or smithereens. In discussing the ontology of holes, Dennett points out, "holes are Gibsonian affordances par excellence, put to all manner of uses in the course of staying alive and well in a hostile world" (p. 102), and thus they have a strong claim on thinghood, even though "it is surprisingly hard to say what holes are made of, if anything, what their identity conditions are, whether they are concrete or abstract, how to count them, and so forth" (p. 102). It is but a small step from this point to noting that if we take stances as collections of bodily postures and coping strategies, holes will show up as important and salient entities with straightforward identity conditions from certain stances and will not show up at all from others. Voices, like holes, become salient within various clusters of coping strategies. Cahoots and smithereens and lost sock centers, on the other hand, do not have any obvious home in any thick-bandwidth set of practices of this sort.

Money is an interesting example. Dennett (2013) says that it "seems obvious" that dollars exist. But, he insists, they are *abstract*—they have no particular physical embodiment. "It makes sense to ask whether this particular dime or silver dollar once passed through the hands of Jacqueline Kennedy, but not whether the hundred dollars I just sent to my PayPal account ever did" (p. 104). Are dollars *real*? On this question, Dennett remains stubbornly ambiguous in just the way we saw before; he insists on sticking to rhetorical questions and speaking of what it is "hard to resist" rather than coming down either way. Indeed, it's striking, from a rhetorical point of view, just how insistently and for how long he maintains the lack of resolution:

The concept of abstract dollars is undeniably useful in our everyday affairs, but is utility tantamount to reality? This question sits happily

within the perspective of sophisticated naïve anthropology but is out of place in traditional metaphysics. Consider the reality of centers of gravity or mass. Are they real or not? Some philosophers have said yes, and others no . . . Compare the concept of the center of gravity of, say, an automobile with the concept of Dennett's lost sock center . . . The former is an undeniably useful concept, and so palpably useful that you can almost feel a center of gravity once you know what to feel for . . . But however perceptible or tangible a center of gravity is, it is (one would suppose, as a metaphysician) in exactly the same basic ontological genre as Dennett's lost sock center, and kazillions of other equally pointless abstractions we could invent. Must we fill our world with all that pseudo-Platonic garbage? . . . Is there anything dangerously relativistic in acknowledging that the two images may have their own 'best' ontologies, which cannot be put into graceful registration with each other? (2013, pp. 104–105).

It seems to me that we should want a story that makes money real, and lost sock centers unreal—or at least, money should come out *more* real than lost sock centers. I think we can get that result if we see the limitations of Dennett's claim that money is *abstract*. We do this by again focusing on stances as *embodied performances* of a certain sort. It is true that dollars themselves have no characteristic physical realization. Indeed, the reality of money is sustained only by our coordinated social recognition of it as real; if we all stopped taking it seriously, there wouldn't be an "it" there for us to fail to get right.<sup>2</sup> This might seem to suggest that money is at best only quasi- or instrumentally real. But the catch is, we *can't* just stop treating it as real; our coordinated social recognition of money is material, not abstract. Money is caught up and intertwined almost maximally robustly in a huge number of our concrete practices. How much of it we have determines what we wear and eat, where we live, and so forth. An enormous number of our daily actions are directed toward getting it, calculating out how much of it we have in our pocket, spending it, and so on. Meanwhile, how we interact with and respond to one another, at an intricately embodied level, is shaped in all kinds

---

2. In Slavoj Žižek's (1997) words, money is a "fetish object" sustained by "constitutive misrecognition": "The fact that money enables us to buy things on the market, is not a direct property of the object-money, but results from the structural place of money within the complex structure of socio-economic relations; we do not relate to a certain person as to a 'king' because this person is 'in himself' (on account of his charismatic character or something similar) a king, but because he occupies the place of a king within the set of socio-symbolic relations; etc." (p. 105).



of ways by how much money we perceive one another as having and how much money we have and have had in the past and expect to have in the future. That is, economic class shapes microinteractions and perceptions and expectations in untold ways. None of this is at all abstract.

I claim that money is best understood as real in just the same way as beliefs and desires, which is to say plain old real. Dollars and other economic entities are robustly disclosed by (and only by) our embodied economic practices. We cannot deny the empirical reality of money; too many of our concrete activities depend on it and are shaped by facts about it. At the same time, conversely, unless one takes up the *economic stance*, these entities get no grip and can't possibly be disclosed. A cat, quite literally, cannot *see money*. Not even a brilliant cat. Cats just aren't capable of taking the economic stance. And this stance is *not* best interpreted as having a bunch of standing intellectual beliefs about money or a set of intellectual rules for interpreting economic reality. If this is what it came to, we could just decide to quit or ignore them, and we can't. Our very livelihoods depend on our belief in money and our ability to take the economic stance—this is no hyperbole. An alien outsider might be able to model our economic behavior, up to a point, and identify the tokens that serve as money fairly reliably. But—at least so I claim—the ontology and reality of money in all its subtlety can only show up fully to someone with an embodied, skilled understanding of what it is to need it, have it, use it, recognize when others have it, and so forth.<sup>3</sup>

Dennett does not actually tell us what it is to *take a stance*, and although I have hypothesized that he would not object to my portrayal of stances as embodied coping strategies, he does not discuss their embodiment. This strands him without a clear ontological position on the reality of money and like entities. While he usually wants to say that folk ontology *just is* ontology, he is left without an argument as to why we couldn't just opt out of taking particular stances, which in turn makes their corresponding ontologies feel kind of optional, and hence shady

---

3. Eric Winsberg (2014), in private conversation, points out a possible confusion arising from my discussion. There is a definite sense in which money is “socially constructed” while trees and rocks are not. Were it not for social attitudes and norms, there would be nothing that could in any sense count as money. But my point is not that one must have the economic stance in order to see money *because* it is socially constructed. If anything, my point is almost the converse: *despite* the fact that money is socially constructed, once we are embedded in the economic stance, its reality is as unavoidable and material as that of trees and rocks. And money, like trees and rocks, helps shape the world to be negotiated, even for those creatures who cannot take up a stance from which it appears. Hence the social constitution of a thing does not water down its reality. In contrast, cooties are socially constructed in a different sense; although we (4th graders) have social attitudes about who has them, they place no interesting material constraints on us from any stance, and require no coping skills. Cooties are not real.



in their claim on reality. But once we note the concrete character of stance taking and all the constraints that come with that, this feeling of optionality dissipates.

In defending the idea that competing and overlapping patterns can be equally real, Dennett (1991) writes, “There could be two different, but equally real, patterns discernable in the noisy world. The rival theorists would not even agree on which parts of the world were pattern and which were noise, and yet nothing deeper would settle the issue. The choice of a pattern would indeed be up to the observer, a matter to be decided on idiosyncratic pragmatic grounds” (p. 118). But this voluntaristic and idiosyncratic language is misplaced, on my reading. Which patterns are salient to us is in some sense relative to our pragmatic goals, but these are typically neither idiosyncratic nor up to us. As we move through the world we need to cope and coordinate with objects, situations, and other bodies, and their material reality—including their instantiation of intentional patterns and teleological patterns—tightly constrains how we do this. If we don’t adopt the right stance, we are likely to end up smashing into people and unable to use the equipment around us or feed or shelter ourselves. Again, I don’t take this to be a substantial modification of Dennett’s own view, so much as a version of it that insists on the materiality and non-optionality of the objects revealed by stances.

Dennett’s *usual* line, which I think is right, is that once we see whether a stance is successful in giving us coping strategies for negotiating some system, there is just *no further question to be asked* about whether the objects of those strategies are “real.” Later in the chapter, I will defend this line in detail. I will argue that any attempt to distinguish between levels of reality with differential legitimacy, or between something like strategic and “real” reality, is confused—and hence that Dennett’s own sporadic defenses of “mild realism” or “in between realism” are likewise confused.

Responding to Dennett in “Pattern and Being,” John Haugeland (1997) argues that adopting a stance requires *taking a stand* (p. 284). Once we take the materiality of stances seriously, it seems to me that “stances” and “stands” are pretty much interchangeable expressions. For Haugeland, however, this move brings out the point that stances are not just repertoires of behaviors; they require that we take on various *commitments* to things behaving a certain way, treating it as a problem or a challenge if they do not. To take the intentional stance with someone is not just to read her as having beliefs and desires, but to be committed to doing so, so that their violation of basic principles of rationality shows up as a norm-violation. From the intentional stance, we not only *take* people as having beliefs and desires that are appropriately connected to one another, but we also *hold them* to doing so, and we take it as a *challenge* to our own stance if this doesn’t work. This point is, I think, built into my material reading of stances. On my reading, we simply can’t interpret *taking a stance* as holding a merely intellectual set of attitudes or

engaging in merely theoretical explanatory strategies. Taking a stance involves engaging a set of strategies for wrangling with resistant material things of various sorts, and doing so necessarily involves *trying to implement* these strategies, perhaps in the face of resistance, *being thwarted* by things, and so forth. Indeed, I want to claim that it is only in the face of this sort of commitment-requiring resistance that we ought to consider the entities toward which our stances are directed as *real*. Entities that could not thwart us in this way, or that cannot be held to specific norms of behavior, don't have the right kind of pushback to count as objective.<sup>4</sup>

Interestingly, we don't need to insist that a strategy or stance work *perfectly* for coping with a system in order for us to count its entities and features as real. People need not be impeccably rational belief-desire machines in order for their beliefs and desires to be entities with which we can cope effectively and coordinate our behavior. Indeed, it is partly because entities can resist our coping strategies and break down in various ways that we take those strategies as responsive to real features of them.

Indeed, we keep discovering that intentional psychology is more partial and imperfect than we thought. It turns out that even when people experience themselves as making choices based on their beliefs and desires and executing their intentions, in fact their behavior is driven to a startling degree by subpersonal pressures. Implicit biases and contextual cues drive our reactions, expectations, and choices to a dramatic degree, and this is often cut free from the phenomenology of action.<sup>5</sup> R. Scott Bakker (2012) argues against the legitimacy of the intentional stance, by way of a discussion of the recent development of "neuro-marketing" and other subpersonal control techniques; he proclaims, "Make no mistake, the 'Age of the Subpersonal is upon us.'" But such subpersonal phenomena do not undercut the intentional stance; they merely showcase its finitude. The intentional stance is indeed a much more approximate tool than we presumed it was in the 1980s and 1990s, and it turns out that our first-personal uses of it on ourselves are not interestingly more precise than are third-person uses. But coping is by its nature an approximate, good-enough kind of activity.

I have claimed that there is no more to the question of whether entities are real than whether they behave in counterfactually stable and predictable,

---

4. This is one decent reading of Kant's distinction between objective perceptions and subjective impressions in the First Critique, and I intend it to be a fairly direct gloss on Haugeland's account of objectivity in "Truth and Rule Following," in *Having Thought* (1997, Ch. 13).

5. There is far too voluminous a literature on implicit bias and related phenomena to cite here, at this point. For philosophically rich discussions of the significance of these phenomena, see, for instance, Huebner (2009, 2016).

norm-governed, and resistant ways as we cope with them from within the stance from which they show up. Since I am now claiming that entities will typically do all this only approximately and imperfectly, it seems that I am committed to saying that being real or objective is not an all-or-nothing thing—that an entity can be roughly and approximately real, *even though* there is no separate standard of the “really real” or of “strong realism” that it is failing to meet. This is a bullet I am pleased to bite. I think the result seems odd only if we impose an ultimately incoherent standard of uber reality or stance-neutral reality. It is to deflating such a standard that I now turn.

### 3. There Is No Neutral Stance

Not infrequently, Dennett seems tempted to treat the physical stance as some sort of fundamental, neutral stance; he talks about how well various entities can fit into the world of “atoms and molecules,” calls patterns “arrangements of atoms and molecules,” and sometimes associates *literal* truth with truths about such physical entities.<sup>6</sup> In “Kinds of Things: Towards a Bestiary of the Manifest Image,” Dennett (2013) draws on Sellars’s distinction between the scientific and manifest image right in the title—a distinction founded on the idea that the physical stance has some sort of distinctive neutrality and freedom from human constitutive engagement.

But there is no stance outside of all stances; there is no neutral stance. The entities of physics themselves can show up only from a distinctive stance, using special tools and resources. When we take seriously the embodied character of stances it becomes clear that one always has to be in *some stance or other*, to the extent that one is functioning as an agent at all. This point becomes fairly obvious, verging on trivial, once we explicitly conceptualize stances as embodied postures and strategies. There is no way for us to engage with the world without doing so with our bodies, and in doing so we need to hold and use our bodies *in some way or other*. If we think of stances as intellectualized sets of explanatory goals and heuristics, we might think it is possible to be in a state free of any such set. But once we think of a stance as an embodied posture, it becomes manifest that one must always be in one or another.

Furthermore, whatever stance we are in, it will reveal some sorts of things to us and hide others, it will create expectations, it will ready us for some sorts of activities and engagements and destabilize us for others, and it will enable some sorts of coping and preclude others. Taking the purely physical stance renders

---

6. See, for instance, Dennett (1987, 2013).

us utterly incapable of coordinating with other acting human beings. Taking the intentional stance quickly becomes frustrating if it turns out we are talking to an automated computer on the “customer-service” line. When the cluster of strategies that make up a stance prove unhelpful, we sometimes switch stances. But there is no uber stance that we can take, from which we can assess which stance properly matches some stance-independent reality.

Stated as such this point seems fairly evident, but I think its implications are important. Most immediately, it makes us reassess the question, Do the entities that are salient from the intentional stance—propositional attitudes—*really, really* exist? There are two possible readings of this question. On one reading the question is, how robustly do they exist? How richly resistant and entangled in our coping skills are they? Are they more like trees and money, or like cahoots and smithereens? Here the answer seems to be: quite robustly indeed. It’s very hard or impossible to negotiate the world around us unless we appeal to intentions and other such mental states, unless we are hermits. According to the other reading, the question is, are they literally, non-scare-quotes real like atoms and trees, or are they just some kind of as-if real that is stance-dependent in a special way? This second question, I claim, is misguided. It presumes that there is some privileged stance outside the intentional stance from which such a question can be asked. From *inside* the intentional stance, the answer will be yes, of course these things really exist. They are precisely what we are coping with and negotiating. From any other stance, the entities and patterns distinctive of the intentional stance will either not show up at all, or they will show up as derivative or instrumental. But that is almost definitionally what we would expect. And there is no third stanceless posture we can take from which to assess the “real” reality of the things that show up as real within a stance.

From within the intentional stance, we can perfectly well inquire into the reality of any particular propositional attitude; this will be an empirical question. And the stance works well enough, enough of the time, to have earned its bona fides as a stance we *can* take. But the broader question of whether its entities are “really real” or “literally real” seems to presume that there is some stance from which we can compare what shows up within a stance to “reality,” without this just being another stance-based strategy for coping. To put the point another way: Ontology, as I am portraying it, is inherently a within-stance activity. To do ontology is just to make explicit what the presuppositions are about how things must be and what needs to exist in order for a set of coping strategies that form a stance to work at all.

If we get behind the idea that there is no such thing as a stance-independent stance on what is real, or what is “really real,” we need not abandon first-order debates over the reality of various things, including debates over whether some

kinds of entities or patterns reduce to others. These are perfectly legitimate debates, but they are *first-order* debates that appeal directly to evidence and anomalies that show up as we cope with things, rather than metadebates that transcend any stance. Importantly, you can agree with me that reality is a stance-dependent notion and that there is no uber-stance, and still disagree with me wildly about what there is. I defend a highly permissive ontology in which many sorts of things are real. But this position is severable from my rejection of a trans-stance perspective. Someone else might agree with me about the metapoint and yet defend a very sparse ontology, perhaps arguing that most of our folk ontology is elliptical and/or incoherent and that only scientific ontology holds up under (first-order) scrutiny. I will return to this point subsequently.

Here is a possible objection to my argument so far: Isn't my claim that stances are *really* sets of embodied strategies for coping itself a trans-stance claim? In doing philosophy, am I not coming down on what the *best way* to think about them is? Similarly, if I defend a profligate or sparse ontology, am I not making a metaclaim about stances and which are legitimate? I think this objection is tempting but sophistical. Philosophy has its own stance, or set of stances perhaps. The notions of "stances" and "ontologies" are themselves notions that have their homes within philosophical stances. Remember my anecdote about my ex-husband; while he named stances as he employed them, this was because he was a philosopher. Most people engage with objects that show up within stances but do not take stances themselves as entities; philosophers do. Saying this does not give philosophy some special priority or trans-stance status. Philosophical entities like ontologies and stances are rightly judged real to the extent they stand up to scrutiny using the tools of philosophy. In this way, they are analogous to all other entities that show up from particular stances.

Now, entities and patterns that are available from one stance are typically not from others. Hence an attempt to explain *why* a stance works or *what kind of reality* its entities have, in terms *foreign* to that stance, will routinely fail to capture what is available from inside the stance. So, for instance, there is a long and vigorous tradition in philosophy of trying to explain how responsiveness to the kinds of norms that get intentional psychology off the ground can begin. Dennett and Robert Brandom, for example, debate where normatively governed intentionality comes from. Brandom says that the origin is social, whereas Dennett claims it is evolutionary. Of Brandom, he writes, "On this line, only communities, not individuals, can be interpreted as having original intentionality. And membership in a (linguistic) community, while in some sense optional for us, carries with it a commitment to the conceptual norms established and constituted by that community. I, in contrast, have claimed that our own intentionality is just as derived as that of our shopping lists and other meaningful artifacts, and derives

from none other than Mother Nature, the process of evolution by natural selection” (Dennett, 2006, p. 61). But what always feels unsatisfying about these sorts of stories is that however we tell them, the resulting story seems to capture a *different* kind of thing than what we were trying to capture, even if it’s extensionally identical. No evolutionary story and no story about the origins of us beating one another with sticks will ever strike us as capturing the special normative character of rationality.

I propose that this is not because of some sort of magical irreducibility of the mental, but because such stories are doomed to fail from the start—they bring us into a stance other than the one we are trying to explain. They may be perfectly good explanations of co-extensional phenomena. But if I am right that adopting a stance is a matter of taking up a complicated posture that shapes salencies, expectations, possibilities for coping, and the like, then bringing ourselves outside of the stance and entering a different one will make the distinctive character of what we are trying to explain disappear.

This, I think, is why all attempts to “naturalize normativity” seem to miss the point. None of the quasi-historical, quasi-scientistic, or quasi-reductionist stories about why the intentional stance works *themselves* employ the intentional stance; they use different explanatory strategies, appeal to different sorts of covering laws, and direct attention to different sorts of entities and patterns (such as patterns of social coordination, adaption, or neural correlations and their corresponding entities). The *project* of giving an historical or reductionist story is simply not a project that one takes up the intentional stance in order to complete. This doesn’t mean that such stories are false, necessarily. But it does mean that the distinctive character of the intentional cannot show up within them.

This seems mysterious only if we think of stances as collections of intellectual attitudes. Opportunities to punch, requirements to block, corners to escape to and the like only show up properly from a boxing stance; they are not visible from, say, a sprinter’s stance, not to mention a medical clinician’s stance or a physicist’s stance. Once we think of stances as whole-body, largely implicit skill sets and postures, it is not at all surprising or magical that switching makes their entities disappear.

In “Philosophy and the Scientific Image of Man,” Sellars (1962) insists that things that operate within a framework cannot be used to support that framework (pp. 27–28). But this seems wrong to me, and surprising coming from Sellars. The only on-point support that a framework or stance can have is its internal ongoing success as a set of coping strategies. Explanations of why the entities in a stance are real, or why the stance as a whole works, cannot be inter-stance questions if we want to hold onto the distinctive character of the stance; they need to be intra-stance questions. This has the interesting consequence that

stances cannot find or explain their own boundaries. Our *transition* from nonintentional beings to intentional ones is not one that will be explicable from within the intentional stance—as we move to more and more rudimentary systems, the intentional strategy just gets less and less helpful and eventually peters out, much as the visual field does not encompass its own boundaries (Wittgenstein, 1990, *Tractatus* 5.633).<sup>7</sup>

#### 4. But Are the Entities Literally Real?

I have insisted that there is no interstance notion of reality to appeal to, that the real just is that which shows up as resistant and counterfactually norm-governed from within an embodied set of coping strategies. I think this insistence is likely to be met with a stubborn resistance of the following form: But are these entities *literally real*? Or is intentional language just a *way* of characterizing a reality that is *literally* just physical, or biological, or whatever? Remember, Dennett himself casts the problem this way sometimes. In discussing the reality of the objects of intentional psychology, we saw him say, “I think beliefs (and some other mental items drawn from folk psychology) are . . . attributed in statements that are *true* only if we exempt them from a *certain familiar standard of literality*” (1987, p. 72).

Whether or not the language of literality is explicitly invoked—and it often is!—I think that the just mentioned resilient intuition depends on the notion. The idea is that we ought to be able somehow to *compare* the claims yielded by the intentional (or whatever) stance to reality itself and make sure that they correspond. In the previous section, I argued that no such interstance comparisons could really be explanatory, and that there is no neutral stance from which to assess stances. In this section and the next, I want to look more rigorously at the idea that we can assess the *literal* existence of the entities toward which a stance is directed, or the literal truth of the claims it yields.

What is it for something to be literally so? If we are being careful, literality can apply (at most) to declarative sentences or claims; to say that they are literally true is to say that their contents or meanings map onto the world in some standard way—that meaning and world *match*—a matching that is somehow faithful or direct, rather than metaphorical, hyperbolic, or distorted in some other way. There is of course a huge literature on how to make sense of such correspondence claims; we needn’t sweat the details for now. But when we say something

---

7. Elsewhere I have argued that explanations of normativity must always involve retrospective constitutive misrecognition: we can explain how a system is normatively responsive only by appealing to how it was *already* caught up in norms. Such explanations are always backward-looking (Kukla, 2000, 2002). The point here is kindred.



is “literally real,” we are speaking elliptically or imprecisely: we mean something like, “The name of this thing refers and claims about this thing can be literally true.” The notion of *literal reality* only makes sense derivatively, in terms of some sort of comparison between meaningful things and bits of the world.<sup>8</sup>

My hypothesis is that our impulse to insist that there must be some interesting question about whether entities that we can cope with are “really real” or “fundamentally real” or “literally real” is grounded in an intuition that the *primary* ontological question we can ask is actually one about *sentences* or *claims* and what kind of truth they can have—in other words, we take the “matching” question as the *basic* ontological question. This is explicit for Paul Boghossian (1990), who writes, “An irrealist conception of a given region of discourse is the view that no real properties answer to the central predicates of the region in question,” and this “invariably” arises in light of a perceived “mismatch between an account of the meaning of the central predicates and a conception of the sorts of property the world may contain” (p. 157). In other words, for him, questions about realism are inevitably questions about *meanings* and whether they *match the world* in the right way. Metaphysical debates are fundamentally semantic debates on this picture. And from that starting point, my equation of the real with that which resists and yields to our performed, embodied coping strategies appropriately can seem deeply unsatisfying: it doesn’t seem to answer the literality question or even offer the materials for doing so; it also leaves a residual worry that we are coping by casting things under some nonliteral, distorted veneer of meanings. Even if taking things as beliefs and desires is *strategically effective*, it still might be that sentences about beliefs don’t literally correspond with entities in the world, on this line.

There are, I think, two related deep problems with this sort of semantic worry. Once we bring them to light, we deflate an entire range of ontological concerns.

- (1) There is no good reason to prioritize the “matching” question. We do not need to turn questions about reality into questions about the meanings of sentences about reality and how they hook back up with it; while such semantic questions may have their place, this is to introduce an extra layer of metaphysical complexity unnecessarily, and to change the topic. We need not grant Boghossian’s claim that first-order debates are “invariably” about

---

8. There is another, perfectly reasonable, informal sense of the term *literally*: It means something like “no, I am not kidding or being tricky or playful here, nor am I lying or exaggerating for effect,” or perhaps better, “my words right now mean what you’d expect them to mean, as long as I am not being weird.” I have no problem with this sense of literality. But it is not the kind that semanticists and metaphysicians are interested in.



a mismatch between meanings and reality. Indeed, this has little to do with how such questions are framed within science and ordinary language, where terms like “meaning,” “truth conditions,” and “denotation” rarely come up. Typically, debates over reality *just are* debates over whether purported entities behave as entities do—are they predictable and manageable, can they resist and surprise us, how do they support counterfactuals? etc. These are the first-order debates that my account has no problem encouraging, as I discussed in the previous section. To switch to a debate about meanings and matching is, in effect, to change the topic.

- (2) And crucially, given my argument so far, it is to change the topic to *one that makes sense only from a specific stance*. The stance from which we ask such semantic questions is one from which the entities that show up are things like propositions, meanings, denotations, references, and truth-values. These things don’t show up from, for instance, the physical stance or the design stance—indeed as philosophers have long recognized, they are rather curious entities, and it takes a very specific kind of intellectual posture for them to show up at all! But this means that it is stance-hopping to ask questions about the *literal* reality of the entities of other stances. To ask them, we have to move to a stance from which such semantic questions make sense, but in doing so we have left behind the stance that we are examining. And as I argued in the previous section, you can’t do this without losing sight of the phenomenon you are trying to interrogate or explain. It is not necessarily illegitimate to ask semantic questions about the *literality of claims about intentional psychology*. But one should not expect doing so to give us insight into *the nature of the entities of intentional psychology themselves*, as this is available only from within the intentional stance.

But, one might counter, propositional attitudes like beliefs have *aboutness*. Unlike tables and chairs and cogs and carburetors, they are things with *meanings*. Hence semantic questions can get a grip here in a way that they maybe cannot if we are inquiring into the nature of biological entities or the entities of physics, say.

But this objection is based on a tempting yet straightforward scope error. Questions about the accuracy of particular intentional states may indeed invoke semantic questions about their meaning and relationship to the world.<sup>9</sup> But questions about the *nature or reality of intentional states themselves* are not questions about their content or literal truth. These questions are neither more nor less

---

9. Although typically during the flow of employing the intentional stance, such questions don’t really emerge very often, and when they do they arguably jolt us out of the stance at least temporarily. I discuss this in the section, “Literality and the Interpretive Stance.”

about semantic questions than are those about the nature of theoretical entities in physics or units of selection. Of course, if we want to, we can look at statements about beliefs and desires and inquire into their literal truth, which will require doing an analysis of their meaning. But here as elsewhere, the assumption that questions about what is real must be translated into questions about the semantics of certain kinds of sentences is unjustified. The fact that the entities under question themselves have meaning is neither here nor there.

To make sure the point is clear, let's look at a few different questions I might ask, to clarify the relationship between questions about literality and questions about realism. I might ask, "Does Ada believe that cats are fuzzy?" This is a question normally answered from within the intentional stance. I look at how Ada acts and what she says. I *could* take it as a question about the content of an utterance of hers, "Cats are fuzzy!" or "Kettir eru loðnar!" Normally, I would do this if I had some question about how she used language. But this is a bit of a jolt out of the normal intentional stance. Typically, from within it, I am not directing my attention toward meanings or matching contents with objects, but rather attending to how people act. I might instead ask, "Is Ada's belief that cats are fuzzy literally true?" This is an odd sort of question, but again, normally it would be answered by directly examining the evidence concerning cats and their fuzziness, unless the issue is specifically that we are not quite sure how to interpret the content of Ada's belief. It seems that questions about literality can emerge from inside the intentional stance, but they tend to involve a kind of a fissure in it rather than part of business as usual. In any case, neither of these questions are ontological questions about the nature of the object of intentional psychology.

If we ask instead, "Does Ada have any beliefs?" or "Are there really such things as beliefs?" then these are questions about the reality of intentional states. But these are not the same questions as "Are some statements about Ada's beliefs (or about beliefs in general) literally true?" The former are questions about the ontology of the mind, and the latter are questions about semantics, and it strikes me as a piece of dogma to think that the two are somehow equivalent. I don't see any reason to double and complicate the ontological issue by asking whether beliefs are not only real, but *literally* or *really* real, nor do I see any way of making sense of those questions except by interpreting them as implicitly or explicitly moving to the semantic level.

## 5. Literality and the Interpretive Stance

While beliefs and desires are specialized entities that show up only from within the intentional stance, this is at least as true for meanings and other semantic

notions—these things don't make up part of the common-sense furniture of the world at all. So from what sort of stance *does* the notion of literality get a grip, exactly? If semantic entities are so weird, do we need them at all? Maybe we should just ditch the notion of literality altogether, on the ground that semantic notions such as meaning, denotation, and truth are not things that we normally need to draw upon as we cope with the world around us. But I have rejected the concept of a stance-free stance, and with it the notion that there is some fundamental objective ontology. So the question is really, is there a productive and useful stance that attends to and makes use of these semantic notions, from which literality is a meaningful standard to which things can be held?

The answer, I think, is that semantic notions including literality show up primarily in what I will call the *interpretive stance*—that is, the stance from which we employ a variety of strategies in order to decode the meaning of what another system is saying or thinking. This can come up when we are translating, or when we hit a snag in our ability to make sense out of what a linguistic being is up to. If our *project* is one of mapping sentences to sentences, or sentences to the world, then as a matter of structural necessity, the ontological toolbox we will use will be a semantic one. This can be a perfectly reasonable toolbox to use even if we do not think that meanings or a truth property are somehow part of a stance-free objective ontology. In fact, the interpretive stance normally kicks in only under when a sort of breakdown occurs: We don't normally question or direct attention to someone's meaning until we become unsure how to map their sentence onto our own beliefs. The primary home of meaning—the place where it becomes something that we need to cope with as an entity—is in a communicative context in which it is contested. In such a context, we can distinguish between the words someone utters and what she means by those words, because it is always possible to misinterpret or misunderstand or simply not grasp the import of the words. Hence it is always sensible to separate the utterance from its meaning and ask if we got the meaning *right*. But that meanings show up in this sort of context does not show that they have some kind of interstance uber reality that can be used to measure the literal reality of the entities of other stances.

Perhaps the most articulate and paradigmatic developer of the interpretive stance is Donald Davidson, and he is also the best example of a philosopher who makes use of the ontology of the semantic in a way that is consonant with my point. Across multiple papers and decades, Davidson developed a picture of interpretive practices, in which we come up with a theory of meaning for a speaker. The key elements of Davidsonian interpretivism, for our purposes, are as follows: The basic unit of meaning is the truth-valuable sentence. We develop a theory of meaning by developing a holistic theory that triangulates between the

speaker's sentences, our own beliefs, and the world, designed to maximize the speaker's correctness and rationality.

To do this, we take semantic notions such as truth and satisfaction, which involve matching bits of language to the world, as primitive. With that move, Davidson explicitly invert's Tarski's order of explanation while using his tools: Tarski wanted to define truth in terms of a set of biconditionals of the form "S' is true if and only if *p*," assuming it as already understood that the right sides of the biconditionals were translations expressing the meanings of the sentences on the left. Davidson (1973) "propose[s] to reverse the direction of explanation: assuming translation, Tarski was able to define truth; the present idea is to take truth as basic and to extract an account of translation or interpretation" (p. 134).

In the interpretive stance, we take up a set of strategies for understanding a speaker and coping with her in specific ways—strategies that involve uncovering meanings by correlating sentences, interpreter beliefs, and the world. The picture is familiar to philosophers. It invokes our capacities to identify assertions and agreements, employ the principle of charity, pick up on how a speaker's attention is directed, and so forth. While this stance is related in interesting ways to the intentional stance, and while there are often times in which one moves quickly back and forth between them or perhaps even employs both, it is importantly distinct. The point of the interpretive stance is to assign meanings to *sentences*—to correlate bits of discourse with either the world or other bits of discourse.

Davidson insists in several different ways on the importance and the coherence of the notion of the literal. For instance, he writes, "Metaphors mean what the words, in their most literal interpretation, mean, and nothing more . . . The central mistake against which I shall be inveighing is the idea that a metaphor has, in addition to its literal sense or meaning, another sense or meaning" (2001, p. 32) He insists that literal truth conditions and meaning can be assigned to sentences, and that this is separate from how context of use allows us to get something more or different out of a metaphor (p. 33). Likewise, in "A Nice Derangement of Epitaphs" (1986), he insists that "nothing shall obliterate" the crucial distinction between literal meaning and speaker meaning.

He is right to insist on the notion of literality within the context of his project. For Davidson, interpreting is a matter of developing a substantive truth theory that maps sentences onto the world. The Tarski sentences, for him, are contingent empirical truths about how this mapping goes. One of the strongest and most important claims he makes about truth is that we should not be fooled by the *apparent* triviality of many of our Tarski sentences, such as "'Snow is white' is true if and only if snow is white." Such a sentence is true only because it happens that our metalanguage and object language correspond, so that the true meaning

of “Snow is white” happens to be that snow is white. But it could have been otherwise; we could have used “grass is green” to mean this, he insists; the T-sentence here expresses an empirical theory (1967, pp. 311–312). It is our substantive truth theory—for Davidson, a top-down interpretive semantic theory—that will tell us whether a metalinguistic phrase maps onto the homophonic object language equivalent. It is only a matter of “convenience,” as Davidson puts it, when our metalanguage and object language are homophonic or homographic. But in such a context, literality, or proper matching of sentences with sentences by way of a substantive notion of truth, *is* the measure of meaning. If the Tarski sentences are substantive empirical claims, then there have to be criteria for getting them right or wrong.<sup>10</sup>

I claim that the ontology of the literal only actually gets a grip within this sort of interpretive project. We need the specific stance of the interpreter in order for the elements of interpretations to show up as the kinds of things we can cope with and comport ourselves toward. The intentional stance may *seem* similar because it involves assigning propositional attitudes to systems, and those propositional attitudes have content. But in fact, focusing on interpreting content and worrying about how it matches up with the world or with other sentences is quite different from what we normally do when we interact with someone as an intentional system.

The interpretive stance, insofar as it depends on a semantic ontology and makes use of the notion of literality, relies on an inflationary picture of truth and semantic facts. But Davidson’s picture of interpretation and its reliance on a robust semantic ontology is not the only game in town. For deflationists about truth, unlike for Davidson, truth is disquotational. T-sentences of the form “‘Snow is white’ is true if and only if snow is white” are not substantive contingent components of a theory of meaning, but trivial conceptual facts, and the truth predicate serves only as bit of logical formalism allowing us to express things like indirect discourse and semantic ascent.<sup>11</sup> On such a picture, matching sentences with sentences or sentences with the world is not a matter of identifying meanings and checking to make sure they are literally identical or mapping

---

10. Interestingly, and appropriately, Davidson himself portrays his interpretive stance as a set of embodied coping strategies; his portrayal resonates with my characterization of stances more generally. In “A Nice Derangement of Epitaphs,” he writes, “We have erased the boundary between knowing a language and knowing our way around the world generally” (1986, p. 107). Interpretive success, he claims, involves deploying passing theories developed on the fly, using “wit, luck, and wisdom” (p. 107).

11. See, for instance, Field (1994); and Kukla and Winsberg (2015).

them onto facts, but rather a matter of finding ways to successfully coordinate our behavior.

There is no such thing as synonymy or lack thereof for a deflationist, but only adequately successful or unsuccessful coordination. Translation is not the kind of thing for which there is a meaningful standard of exact or perfect success. To ask whether two expressions are synonymous is to ask whether they share the same meaning. But if meaning is a thoroughly disquotational notion, and meaning facts are trivial rather than contingent, then there is no more to be said about the meaning of “S,” for any given speaker, than that it is S. If Sarah says “A,” and Joe says “B,” then Sarah’s utterance is true if and only if A (in Sarah’s language, as she presently understands it), while Joe’s utterance is true if and only if B (in Joe’s language, as he presently understands it). There is no metalinguistic stance from within which we can ask whether “A”-for-Sarah means the same thing as “B”-for-Joe, because there is no nontrivial relationship between A and “A’ is true” or between B and “B’ is true” to be probed. For the deflationist, then, the biconditionals do not tell us distinct semantic facts about sentences that can then be compared. And without a robust notion of synonymy as literal, contingent sameness of meaning, there is no such thing as literal translation between languages.<sup>12</sup> Conversely, if it is contingent that “S” is true if and only if S, or that “p” refers to p, then we need an ontology that can make such claims meaningfully true or false, and this will give us notions like synonymy and literality.

Elsewhere, Eric Winsberg and I develop and defend a deflationary picture of truth and semantic facts.<sup>13</sup> My goal here is not to adjudicate between a deflationary and an inflationary picture. Instead, I want to make two points:

First, I’ve argued that semantic entities only show up as things to be coped with from the interpretive stance, and that the interpretive stance itself is most often useful during moments of communication breakdown or challenges. This is a fairly contained domain. Within this domain, the issue between the deflationists and the inflationists is whether things like meanings and reference relations have enough resilient robustness and systematicity to count as real, or whether they are best thought of as instrumental elements of a rough model. This is an empirical question that I do not try to answer here. But remember, earlier I argued that entities that show up from a particular stance can be roughly or approximately real, and that reality is not all-or-nothing. I strongly suspect that

---

12. A similar point is made by Derrida and poststructuralists. Much of this paragraph paraphrases Kukla and Winsberg (2015).

13. Kukla and Winsberg (2015).

the right thing to say is that semantic entities are real-enough-for-government-work in some limited domains and not part of the ontology of most.<sup>14</sup>

Second, the deflationary picture is useful to us regardless of how we settle the above issue, because it allows us to make sense of debates over the reality of some domain without needing to appeal to literality or meaning or a substantive truth property, which means we are not stuck universalizing the interpretive stance and treating its entities as somehow especially fundamental or indispensable. Since the deflationist does not believe in substantive facts about truth, she also does not believe that different sentences can be true in different ways, or that they can have some sort of nonstandard, nonliteral form of truth—or, for that matter, a standard, literal form of truth. Boghossian (1990) criticizes the deflationist for being unable to make sense of the distinction between factive and non-factive discourse. But once we have dispensed with the idea that questions about what exists have to be interpreted as questions about whether bits of discourse are factive—in the sense of having meanings that correspond to the bits of the world they talk about—we can turn this critique on its head: Part of the power of deflationism is that it does not depend on an ontology in which there is such a distinction, at least at the level of semantics. The deflationist can perfectly well distinguish between discourse that helps us navigate the world effectively and coordinate our behavior, and discourse that does not. There is no further question, for her, about whether a sentence has a meaning that matches literally either with the world or with a different sentence's meaning.

Thus, even if there are important uses for the interpretive stance and hence for semantic facts in some specific practical contexts, there is no reason to insist on such an inflationary picture undergirding general inquiries into what is real. Nor, as we saw earlier, should we stance-hop, and import the ontology of one stance into the assessment of another.

## 6. Ontological Questions Are Practical Questions

On the view I have outlined, ontological questions about what is real deflate to practical questions about what sorts of things we cope with from a particular stance. This is not to reduce metaphysical questions to pragmatic epistemic questions; we can of course be *wrong* about what it is we are dealing with. It is rather to deny that there is any *separate question to be asked* about the literal reality of something beyond questions about whether it is there to be coped with. Metaphysical

---

14. This represents a slight shift in my view since Kukla and Winsberg (2015).



questions are not in the first instance semantic questions or questions about truth or reference. To assume that underneath first-order debates are implicit questions about whether objects correspond to meanings is to have already adopted a deeply inflationary stance, and to have inappropriately universalized that stance.

Our intellectual attitudes do not create reality; nor do our richer embodied stances. However, the stances we take allow different features of reality to become salient and detectable. Such stances are indispensable in determining what sorts of existence things show up as having. I take it this was Sellars's (1962) point in "Philosophy and the Scientific Image of Man," when he writes,

A primitive man did not believe that the tree in front of him was a person, in the sense that he thought of it both as a tree and as a person, as I might think that this brick in front of me is a doorstep. If this were so, then when he abandoned the idea that trees were persons, his concept of a tree could remain unchanged, although his beliefs about trees would be changed. The truth is, rather, that originally to be a tree was a way of being a person, as, to use a close analogy, to be a woman is a way of being a person, or to be a triangle is a way of being a plane figure. (p. 10)

Sellars continues by saying that we cannot properly be said to *believe that* things have the kind of being they do—that a triangle is a plane figure, for instance. I take it his point is that taking them as plane figures is built into the stance we take toward triangles, rather than something we discover about them. It is a transcendental feature of them, to use Kantian terminology. Indeed, this Kantian turn is useful for me. What performed stances do is set the transcendental conditions in which certain kinds of things can show up, and they do it in a concrete, embodied way, not through providing purely intellectual frameworks.

One can ask whether someone's sentences about intentional objects (or designed objects, or whatever) are literally true, but this is a question about how to interpret their meaning, and it has its proper home in a specific kind of communicative context. Appealing to literality does not give us a privileged method for doing ontology. One can ask questions about what is real from within any stance, and one can, upon investigation, give fallible but well-supported answers. One will get these answers by employing the body of coping strategies available from that stance. By these standards, I think that the best evidence we have is that beliefs and desires and intentional systems are indeed straightforwardly real. But there is no extra-stance perspective from which to assess the correctness of a stance.



## Works Cited

- Bakker, R. S. (2012). Getting subpersonal: Should Dennett rethink the intentional stance? [Web blog post]. Retrieved from <https://rsbakker.wordpress.com/2012/12/17/getting-subpersonal-should-dennett-rethink-the-intentional-stance/>
- Boghossian, P. (1990). The status of content. *Philosophical Review*, 99, 157–184.
- Davidson, D. (1967). Truth and meaning. *Synthese*, 17, 304–323.
- Davidson, D. (1973). Radical interpretation. *Dialectica*, 27, 313–328.
- Davidson, D. (1986). A nice derangement of epitaphs. In R. Grandy & R. Warner (Eds.), *Philosophical grounds of rationality* (pp. 157–174). Oxford, UK: Oxford University Press.
- Dennett, D. C. (1987). True believers. In Daniel C. Dennett, *The Intentional Stance* (pp. 13–42). Cambridge, MA: MIT Press.
- Dennett, D. C. (1991). Real patterns. *Journal of Philosophy*, 88, 27–51.
- Dennett, D. C. (2006). Evolution of why: Essay on Robert Brandom. *Making It Explicit*. (Unpublished manuscript).
- Dennett, D. C. (2013). Kinds of things: Towards a bestiary of the manifest image. In D. Ross, J. Ladyman, & H. Kincaid (Eds.), *Scientific metaphysics* (pp. 96–107). Oxford, UK: Oxford University Press.
- Field, H. (1994). Deflationary views of meaning and content. *Mind*, 103, 249–285.
- Haugeland, J. (1997). *Having thought*. Cambridge, MA: Harvard University Press.
- Huebner, B. (2009). Troubles with stereotypes for Spinozan minds. *Philosophy of the Social Sciences*, 39, 63–92.
- Huebner, B. (2015). Do emotions play a constitutive role in moral cognition? *Topoi*, 34, 427–440.
- Huebner, B. (2016). Implicit bias, reinforcement learning, and scaffolded moral cognition. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy, Vol. 1: Metaphysics and epistemology* (pp. 47–79). Oxford, UK: Oxford University Press.
- Kukla, R. (2000). Myth, memory and misrecognition in Sellars' "Empiricism and the Philosophy of Mind." *Philosophical Studies*, 101, 161–211.
- Kukla, R. (2002). The ontology and temporality of conscience. *Continental Philosophy Review*, 35, 1–34.
- Kukla, R., & Winsberg, E. (2015). Deflationism, pragmatism, and metaphysics. In S. Gross, N. Tebben, & M. Williams (Eds.), *Meaning without representation* (pp. 1–26). Oxford, UK: Oxford University Press.
- Mirolli, M. (2002). A naturalistic perspective on intentionality: Interview with Daniel Dennett. *Mind and Society*, 3, 1–12.

- Sellars, W. (1962). Philosophy and the scientific image of man. In R. Colodny (Ed.), *Frontiers of science and philosophy* (pp. 35–79). Pittsburgh, PA: University of Pittsburgh Press.
- Wittgenstein, L. (1990). *Tractatus Logico-Philosophicus* (C. K. Ogden, Trans.) London: Routledge. Original work published in 1921.
- Žižek, S. (1997). *The plague of fantasies*. London: Verso.

# 1.2 REFLECTIONS ON REBECCA KUKLA

Daniel C. Dennett

I will treasure the image of Rebecca Kukla's former husband announcing brief cancellations of interpersonal relations with her by shouting "Design stance!" whenever a hunger-induced bout of subrationality temporarily disqualified her for treatment from the intentional stance. More importantly, I endorse the larger claims she makes with this example. The primary role of the intentional stance is not the dry intellectual role of an *option*, deciding which framework will work best to couch one's attempts at explanation and prediction of complex behavior; the primary role is to enable people to cope with each other's differences at a level that privileges our rational capacities, permitting us human beings to be persons, agents with agendas and priorities that deserve respect to the greatest extent possible in our ongoing efforts to coordinate our activities. And adopting the intentional stance is not typically voluntary but irresistible, not a shift in intellectual focus but somehow a more bodily matter of, well, *stance*: the readiness to refrain from some physical actions—pushing and shoving, restraining, tackling—in favor of less physical alternatives—requesting, persuading, cajoling, reassuring. This is why patients in surgery are draped to expose only the part that will be operated on. It makes it easier for the surgical team to set aside the intentional stance.

But another role for the intentional stance is indeed an optional, if persistently tempting, strategy, a bit of intellectual jujitsu that dramatically simplifies and renders tractable (at a cost) the complexities of the behavior of plants and animals, robots and chess-playing computers, and evolutionary design processes, to name the most important. (I was inspired to write about the intentional stance when I first began hanging out with artificial intelligence (AI) programmers in the sixties. I never heard anyone bother to explain or attempt to justify their use of intentional language when discussing their programs; it came naturally, with no scare quotes or raised eyebrows, so it

certainly *seemed* to be literal usage.) What I have all along wanted to draw attention to is the remarkable fact that we human beings have taken our evolution-designed, “instinctual” thinking tool, the intentional stance (aka folk psychology or theory of mind) and adapted it for use in discerning the explanatory patterns in evolution and other design processes. What works to explain the intelligent designer (or reverse engineer) also works for the Blind Watchmaker, and for the same reason.

The cost of this bold extension into phenomena that are not intuitively “mental” is that one way or another we have to countenance something like thermostats with beliefs, bacteria with propositional attitudes, Mother Nature with intentions in spite of having no foresight. There are various ways of providing this elasticity. Most of the philosophers arrayed against me over the years have wanted to defend criteria for “real belief” that rule out all but the “higher” animals, explaining the undeniable utility of the intentional stance in a much wider domain as a matter of convenient metaphor, or short-hand, a mere *façon de parler* in one way or another, not “real belief.” In the early years, and even as late as *The Intentional Stance* (1987) and *Kinds of Minds* (1996), I was reluctant to abandon entirely the Quinian quest to define intentional contexts logically, in terms of referential opacity and the like. I was still a member of the Propositional Attitude Task Force, but growing restive. As it became more and more obvious to me that I wasn’t putting these tenuous logical insights to work beyond using them as a quick-and-dirty litmus test, a rule of thumb for “intentionality” that didn’t demand either minds or brains from the outset, I quietly dropped them. I guess they survive as historical curiosities: notice how Quine’s “intensional” and Brentano’s “intentional” serve together to limn a fairly intuitive category of “the mental,” and for a good reason: mentalistic talk is always *interpretation*.<sup>1</sup>

Kukla surveys my various forays into ontology, noting that “it is but a small step” from my discussion of Gibsonian affordances, to her nicely articulated position, which countenances degrees of reality without apology. Indeed, it is a small step, most of which, at least, I have just taken in my new book (*From Bacteria to Bach and Back: The Evolution of Minds*, 2017—hereafter *BBB*), where affordances are held to be the best examples of real things. As she puts it, “Voices, like holes, become salient within various clusters of coping strategies. Cahoots and smithereens and lost sock centers, on the other hand, do not have any obvious

---

1. David Haig (unpublished) develops a strikingly simple and bold account of the relationship between information and meaning, in which information is a difference that makes a difference, and the difference it makes is always an interpretation by an interpreter, an account that aspires to span the cases from DNA and RNA to plant and animal signaling to the most refined and sophisticated communicative exchanges between language users.

home in any thick-bandwidth set of practices of this sort” (p. 11). I am tempted by this view, by its practical, metaphysics-deflating attitude to what reality might come down to, but I do see a few problems. Santa Claus, for example, is much realer than, say, Danny Riegel, a minor character in Richard Powers’s novel, *The Echo Maker* (2006). People *engage* with Santa Claus, shaping their children’s early lives to include many close encounters (of sorts) with him, but it seems to me that no matter how richly Santa Claus might come to dominate our lives in the future, he could never cross the threshold and become as real as—let alone realer than—the table on which my coffee cup is sitting. Or, to underline the obvious, I—for one—couldn’t countenance an ontology where God counts as real just because He is so entangled with the coping episodes of so many lives.

This problem may admit of a good work-around, if, for instance, we can find ways of distinguishing two (or more?) varieties of fiction: the fiction of novels, of ancient myths and religions, and of hoaxes (and cooties—a great example) on one end of the spectrum, and the “fiction” of colors, centers of gravity, and bitcoin, on the other. It’s an enticing project, it seems to me, but I can’t advance it yet (though I do make some novel suggestions in *BBB*), where I also treat at length a variety of “things” that are indispensable features of biology: the *free-floating rationales* that explain the existence of all the good design-without-an-intelligent-designer. Most of the reasons that have ever existed on this planet have been unrepresented, unappreciated reasons. We use the intentional stance, applied to Mother Nature, in effect, to discern and individuate these reasons, while insisting that these reasons are not the creations of reasoners. While in some regards, these are paradigmatic “explanatory fictions” alongside parallelograms of forces and centers of gravity, they are also *discoverable*, and hence *mind-independent* features of the world. I do not want the distinction between beliefs and free-floating-rationales to be blurred into nonexistence, largely because I want to show that many of our everyday folk-psychological “explanations” of our own behavior given in terms of our beliefs and our comprehension thereof are distorted by our self-image as comprehenders. Butterflies don’t have to understand the rationale of the eyespots on their wings, and cuckoo chicks don’t have to understand the rationale of their siblingcides, but also, we human beings don’t have to understand the rationales of many of our cleverest ploys, arrangements, and institutions. Part of the simplifying strength of the intentional stance is that it typically ignores the issue of whether the rationale of the brilliant chess move was represented by the player. We human reasoners do sometimes represent our reasons to ourselves before deciding what to do, but nowhere near as often as we imagine that we do. But then, it seems to me, the practical ontology of the intentional stance is a mixed bag of concrete psychological states (soon to be identified with “working parts” of the brain, and granted standing in the ontology of cognitive neuroscience) and

abstracta that play the role of idealizations of competences and foibles that can be seen as diagnostic or as part of the “specs” of a particular agent (along the lines of my quotation of Rich Greenblatt’s immortal diagnosis of an early chess-playing program: “it thinks it should get its queen out early.”)

Kukla’s final attempt to douse the embers of stance-neutral metaphysics is a ringing declaration that while Quine’s “semantic ascent” is always possible, and often valuable, it is not obligatory:

There is no good reason to prioritize the “matching” question. We do not need to turn questions about reality into questions about the meanings of sentences about reality and how they hook back up with it; while such semantic questions may have their place, this is to introduce an extra layer of metaphysical complexity unnecessarily, and to change the topic. (p. 21)

I find this a refreshing insight, but I daresay it will appear to many as a cop-out. I am persuaded by Kukla’s discussion in its favor, however, and I would add that I hope that it is true, except that, as her masterful discussion of Davidson shows, to make such a claim would be to lapse back into the very stance-neutral myth that she is undermining here.

## Works Cited

- Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.
- Dennett, D. C. (1996). *Kinds of minds: Toward an understanding of consciousness*. New York, NY: Basic Books.
- Dennett, D. C. (2017). *From bacteria to Bach and back*. New York, NY: W.W. Norton.
- Haig, D. (unpublished). Making sense: information interpreted as meaning. Retrieved from <http://philsci-archive.pitt.edu/id/eprint/13287>
- Powers, R. (2006). *The echo maker: A novel*. New York, NY: Macmillan.

## 2.1 THE MANY ROLES OF THE INTENTIONAL STANCE

Tadeusz Zawidzki

Of the many conceptual innovations in philosophy of cognitive science for which we have Daniel Dennett to thank, none is more central to his agenda, or more influential, than the *intentional stance* (hereafter, IS; 1987). However, both Dennett's critics and admirers have largely failed to appreciate that the uses to which Dennett puts this idea are multifarious, and the connections between them complex. In various places, Dennett argues that IS can be used as (a) an analysis of our mature *concepts* of the mind and mental states, (b) an accurate model of *quotidian interpretation*, (c) an account of *the nature* of the mind and mental states, (d) the foundation for a *naturalistic understanding* of all mental phenomena, and (e) an important component of sound *methodology in cognitive science*. Although these roles are related, it is not clear that IS can play all of them. Dennett himself is ambivalent about this. While there is textual evidence that he envisions a central role for IS in all five explanatory projects, other passages suggest significant hedges to this agenda. So, one goal of this chapter is to provoke some clarification of this issue.

But there is more at stake than mere exegetical tidiness. Many of the most persuasive critiques of IS rely on a conflation of these five explanatory projects. For example, it is often assumed that because IS fails as an analysis of our mature concepts of the mind and mental states, it cannot constitute an accurate model of quotidian interpretation or a plausible account of the nature of the mind and mental states. And, indeed, it is not obvious how IS can retain its utility to some of these explanatory projects without playing an appropriate role in all of them. For example, how could it constitute an accurate model of quotidian interpretation without also constituting an adequate analysis of our mature concepts of the mind and mental states? And how could

it be any guide to the nature of the mind and mental states, without succeeding at both of these tasks? The main goal of this chapter is to answer these questions. In particular, I want to suggest that, although there are good reasons to hold that IS fails as an analysis of our mature concepts of the mind and mental states, this has *no* implications for its role in the other four projects. In fact, I argue that it is precisely *because* IS fails at the conceptual/analytic task that it remains viable as a model of quotidian interpretation, an account of the nature of the mind and mental states, a strategy for naturalizing the mental, and an important component of sound methodology in cognitive science.

In section 1, I briefly characterize IS and how it might contribute to the five projects identified above. I also provide textual evidence that Dennett intends it to play important roles in all five projects. In section 2, I discuss general motivations for favoring a univocal approach to these five projects, and then outline some possibilities that undermine these motivations. In particular, I suggest that our mature concepts of the mind and mental states might play little role in successful, quotidian, behavioral prediction, serving instead largely justificatory or forensic roles. Section 3 then surveys empirical evidence favoring this perspective. Next, section 4 recounts well-known and persuasive reasons against using IS as an analysis of our mature concepts of the mind and mental states. However, in section 5, I argue that these difficulties leave Dennett's overall project largely intact: IS remains an invaluable tool in philosophy of mind and cognitive science, once its roles in the five projects are properly reconceptualized.

## 1. IS as an Answer to Five Distinct Questions

To adopt IS is, according to Dennett, to treat a system as practically rational. That is, it is to predict and explain its behavior on the assumption that it has (a) appropriate goals, (b) access to information relevant to these goals, and (c) selects the most rational behavior relative to these goals and information—that is, the most efficient means of realizing its goals given the information to which it has access. For example, to treat a chess-playing computer program against which you are playing as an intentional system is to assume that its goal is to checkmate you, that it has access to information about the current position of chess pieces on the board, the rules of chess, and chess strategy, and that it will select the move most likely to increase its chances of checkmating you given this information (Dennett, 1978). Whatever one's view of various controversies in the philosophy of mind, one must admit that it is possible to adopt IS toward physical systems, that we often seem to do so, and that doing so allows us to secure significant predictive dividends.



Given that we all agree that many physical systems are intentional systems, what more, if anything, is there to being a full-fledged mind, with all of the kinds of mental states that have traditionally caused so much trouble for philosophers? What more is there to being a conscious, free, fully intentional agent, like a human being, over and above being an intentional system? Dennett has, over the course of his career, marshaled considerable conceptual ingenuity in defense of a simple answer to this question: nothing. There is nothing more to being a full-fledged mind, in every sense that philosophers have ever worried about, than being a certain kind of intentional system.

The question of whether there is anything more to being a mind than being an intentional system admits of at least five different interpretations:

1. It may be a question about our *mature concepts* of minds and mental states. For example, when a competent speaker of English asserts that an agent believes some proposition P, the concept she expresses may be analyzable in terms of IS; she may *mean* that the agent's behavior can be usefully predicted on the assumption that the agent uses the information that it is raining to come to practically rational decisions about how to pursue its goals.
2. It may be a question about how human beings *interpret and predict* the behavior of other agents *in quotidian life*. For example, when we predict the behavior of an agent that is sensitive to the information that P, we adopt IS toward it, and assume that it will behave in practically rational ways given its goals and the information that P.
3. It may be a question about the *nature* of minds and mental states. For example, believing that P may involve nothing more than being usefully predictable on the assumption that one uses the information that P to come to practically rational decisions about how to pursue one's goals.
4. It may be a question about what minds and mental states must be, on the assumption that only natural properties and objects exist: *Insofar as minds are natural objects*, they just are intentional systems.
5. It may be a question about how *science* ought to investigate behavioral phenomena: When confronted with the task of explaining some behavioral phenomenon, *a mature cognitive science should adopt IS toward it*, treating the system under investigation as if it had access to information and goals in light of which the behavioral phenomenon counts as practically rational.

Some of Dennett's many discussions of IS suggest that he intends it as central to adequately answering all five of these questions about the mind.

Regarding the first and second questions, Dennett writes, "Intentional systems theory is in the first place an *analysis of the meanings of such everyday 'mentalistic'*

*terms* [emphasis added] as ‘believe,’ ‘desire,’ ‘expect,’ ‘decide,’ and ‘intend,’ the terms of ‘folk psychology’ . . . that we *use to interpret, explain, and predict the behavior* [emphasis added] of other human beings, animals, some artifacts such as robots and computers, and indeed ourselves” (2009, p. 339). Regarding the third question, Dennett’s position seems similarly unambiguous: “What it is to be a true believer is to be an intentional system, a system whose behavior is reliably and voluminously predictable via the intentional strategy” (1987, p. 15). Given Dennett’s naturalist agenda, this likewise implies an affirmative answer to the fourth question. Since there are no non-natural objects and properties, if IS provides an adequate account of the nature of the mind and mental states then this account must be sufficient to naturalize them. Finally, regarding the fifth question, Dennett thinks cognitive science requires a regimented, formal version of IS, which he terms “Intentional System Theory,” in which to formulate specifications of cognitive competences (1987, pp. 58–60).

Despite this clear endorsement of IS as central to adequately answering all five questions, there are also passages in which Dennett hedges on this claim. For example, regarding the first question, Dennett (1987) admits that folk psychological concepts are mongrel concepts, of which it is unlikely that IS or any other well-defined notion can provide an adequate analysis (pp. 55–57). Regarding the second question, Dennett concedes that IS is not the only tool we use in quotidian interpretation (pp. 53–54), and admits that the notion of rationality which it employs is a vague, “general-purpose term of cognitive approval” (p. 97). Regarding the third question, Dennett (1991) stresses that IS gives access only to “real patterns” in observable behavior that are invisible without it, not an adequate account of the mechanisms that are causally responsible for behavior. This implies a correlatively hedged answer to the fourth question: IS helps naturalize the mind and mental states only because it shows how and why these concepts can help track real patterns in observable behavior. Finally, regarding the fifth question, Dennett stresses that IS cannot be *reduced* to subpersonal psychology. Quite the contrary, on Dennett’s view, IS plays an entirely different role from subpersonal psychology in cognitive scientific explanations; that is its whole point (1987, p. 59). Inspired by Marr’s (1982) three-level conception of cognitivist explanation, Dennett argues that IS descriptions of cognitive systems serve to specify the cognitive competences that are the targets of algorithmic explanations—that is, according to Dennett, cognitive science needs IS to specify *what* the mind/brain accomplishes, not *how*.

How should we interpret such ambivalence? Perhaps it is inevitable for any theorist who thinks he is what “you get when you cross a Quine with a Ryle” (Dennett, 1994, p. 365). While Ryle’s project focused on the analysis of our mature folk psychological concepts, Quine’s goal was to clear the ground for a

scientific approach to the mind. It is not clear that a single notion like IS can play the central role in both of these projects.

Given these ambivalent commitments, it is worth asking what happens to Dennett's agenda if there is *no one* notion that can play all five roles. On the face of it, these five roles seem independent of each other. For example, one can suspect that our mature concepts of minds and mental states are misleading guides to their natures, in the way that our mature concept of water, at least prior to the 18th century, was a misleading guide to its nature. Also, one can doubt that the natures of minds and mental states are relevant to their naturalization, if one believes that minds and mental states, if they exist, must be supernatural entities, radically unlike anything countenanced by natural science. And, even if one believes minds and mental states are unmysterious natural objects or properties, it may be that our concepts of them have little role to play in the methodology of successful cognitive science. Even if economic agents are just configurations of fundamental particles, this way of conceiving of them is not methodologically useful for economics. Similarly, the conceptual tools that cognitive scientists find useful for explaining behavioral phenomena may be radically unlike our mature concepts of minds and mental states, or even what minds and mental states actually are, be they natural or supernatural. It may even be that our mature concepts of minds and mental states have little to do with explaining quotidian interpretation. Perhaps we interpret and predict each other's behavior in everyday life using completely different cognitive resources, for example, subtle behavioral generalizations.

Thus, it is far from obvious that the five questions raised here admit of the same answer, or even related answers. I will argue that there are good reasons to think that correct answers to some of these questions are independent of correct answers to the others. Before turning to this, however, it is important to make explicit some assumptions that *could* ground Dennett's claims that IS is central to answering all five of these questions. This is because, even if Dennett does not fully endorse IS as a univocal answer to these questions and may reject (some of) these assumptions, they are independently plausible, and remarkably widespread among philosophers of mind. After discussing these assumptions, I raise some reasons for doubting them.

## 2. Reasons for and Against Answering the Five Questions Univocally

Why might someone think that central questions about (a) the content of our mature concepts of minds and mental states, (b) the basis for the quotidian prediction and interpretation of behavior, (c) the nature of minds and mental states,

(d) the naturalization of minds and mental states, and (e) the methodology of cognitive science all admit of the same answer? Suppose one believes the following: Human beings are very good at predicting each other's behavior and the behavior of nonhuman agents. Furthermore, this competence relies essentially on the application of mature concepts of minds and mental states, that is, we succeed at predicting each other by inferring each other's mental states from observable behavior. It is hard to see how such success could be possible if our mature concepts of minds and mental states were not largely *accurate* depictions of the *natures* of minds and mental states. Since these can consist in nothing more than natural objects and properties drawn from the ontology of science, our mature concepts of minds and mental states must be naturalistic. Finally, given that the goal of cognitive science is to better understand these natural mental phenomena, our mature concepts of minds and mental states have important roles to play in cognitive science.

If Dennett holds some version of these background assumptions, then it is not surprising that he assumes IS can and should play all five roles. In his own words, IS is "in the first place an analysis of the meanings of . . . everyday 'mentalist' terms" (2009, p. 339). Because we use the concepts expressed by these terms in quotidian interpretation and behavioral prediction, IS can double as a model of this cognitive capacity. Since we are very good at predicting other agents, it is safe to assume that IS gets something importantly right about the nature of the mind and mental states. Since, if we are naturalists, the nature of the mind and mental states must consist in elements drawn from the ontology of science, it is safe to assume that IS, in being right about their nature, can help explain how the mind and mental states can be natural phenomena. Finally, since cognitive science is likewise in the business of gaining an accurate understanding of the nature of the mind and mental states, it is safe to assume that IS has an important role to play in it.

These background assumptions are so widespread that theorists with virtually nothing else in common with Dennett share them, and put them to entirely parallel uses. Consider Fodor (1987). He disagrees with Dennett about the analysis of our mature concepts of the mind and mental states, about what quotidian interpretation consists in, about the nature of the mind and mental states, about how to naturalize the mind and mental states, and about proper methodology for cognitive science. For Fodor, our mature concepts of the mind and mental states are concepts of concrete, unobservable, internal causes of behavior, which do not necessarily bear rational relations to each other or to behavior. Quotidian interpretation involves inferring these internal causes and then making behavioral predictions on the basis of such inferences. Such inferences are typically correct; that is, the mind really is populated by concrete states that cause behavior, with

the properties encoded in our mature concepts of mental states. These properties can be understood in terms of natural relations, for example, computational relations among mental states, and nomic relations between mental states and their worldly contents. And these are precisely the commitments of the only viable empirical paradigms in cognitive science. Yet despite these massive differences with Dennett, Fodor makes the same background assumptions, and as a consequence, like Dennett, he defends a univocal response to the five explananda identified in the introduction.<sup>1</sup>

Suppose, however, that it turned out that our mature concepts of the mind and mental states play only a very minor role in most successful quotidian behavioral prediction. There would then be little reason to suppose that the analysis of these concepts should play an important role in explaining quotidian interpretation. Furthermore, there would be little reason to think that this analysis constitutes an accurate picture of the nature of the mind and mental states, given that our mature concepts of them play little role in successful quotidian prediction of the behavior they cause; that is, if our only grounds for thinking that these concepts are accurate representations of mental states is that they support successful quotidian prediction, if the latter claim were false, then we would no longer have grounds for thinking this. Suppose, furthermore, that there was strong evidence that our mature concepts of the mind and mental states are more important to justifying behavior than to predicting it. There would then be little reason to expect them to play an important role in explaining how the mind can be a natural phenomenon. Much as the concept of a person has proven notoriously hard to naturalize, due, arguably, to its primarily “forensic” role (Shoemaker, 2014), if our mature concepts of the mind and mental states functioned primarily as components of public justifications of behavior, they might likewise seem mysterious from a naturalistic perspective. Finally, given that cognitive science seeks to understand the natural causes of behavior, there would seem to be little role for our mature concepts of the mind and mental states to play in this project if they were primarily justificatory constructs.

As I show in section 3, there is empirical evidence that our mature concepts of the mind and mental states play primarily justificatory roles instead of causal-explanatory or predictive roles. This suggests that reasons for thinking IS is an appropriate analysis of our mature concepts of the mind and mental states are not *ipso facto* reasons for thinking that it is an accurate model

---

1. Kathleen Akins (1996) is an exception to this univocal strategy. She explicitly denies that our common-sense assumptions about mental states can double as adequate accounts of their nature, relative to naturalization projects, and cognitive science.

of quotidian behavioral *prediction*. Since successful quotidian behavioral prediction is the main reason for thinking that IS is a good guide to the nature of the mind and mental states, an appropriate naturalization strategy, and an important component of cognitive science, this would also put the roles of IS in these latter projects in jeopardy. Fortunately, I do not think that evidence against the centrality of mature concepts of the mind and mental states in quotidian behavioral prediction has these further dire implications for the utility of IS. As I will argue, IS can play important roles in explaining quotidian behavioral prediction, elucidating the nature of the mind and mental states, explaining how these can be natural phenomena, and guiding cognitive science, for reasons that are entirely independent of its role in the analysis of our mature concepts of the mind and mental states. First, I turn to a brief review of the relevant empirical evidence.

### 3. Evidence That Mature Concepts of Mind Play Justificatory Rather Than Predictive Roles

Increasingly, philosophers of mind and developmental and comparative psychologists are expressing skepticism that successful quotidian behavioral prediction requires the application of concepts of mind, and mental states, like the propositional attitudes. For example, Hutto (2008) argues that most quotidian interpretation involves only the attribution of “intentional attitudes,” which are, roughly, relations between whole organisms and worldly states of affairs rather than mental states. Apperly and Butterfill (2009), Apperly (2011), and Butterfill and Apperly (2013) likewise argue that much quotidian behavioral prediction in dynamic, interactive contexts by nonhumans, human infants, and human adults employs a “System 1” mindreading competence that does not employ concepts of minds or mental states like propositional attitudes. In fact, they argue that most established behavioral tests of competence with such concepts actually test only for this lower level competence.

This newfound skepticism about the link between our mature concepts of the mind and mental states and our quotidian competence at predicting behavior is well motivated. There is evidence of sensitivity to behavior caused by mental states like false beliefs in infants as young as seven months of age (Kovacs, Téglás, & Endress, 2010), and in nonhuman species as diverse as scrub jays (Clayton, Dally, & Emery, 2007), ravens (Bugnyar, Reber, & Buckner, 2016), and chimpanzees (Hare, Call, & Tomasello, 2006). It is highly unlikely that such sensitivity is conceptually mediated. Furthermore, the use of mature concepts of mental states, such as belief and desire, for seamless and rapid behavioral prediction in dynamic and interactive quotidian contexts, seems ill advised because of the

tenuous, holistic connections between many kinds of mental states and observable behavior. As has long been appreciated, propositional attitudes lead to behavior only in indefinitely large blocs, rather than individually: any belief or desire is compatible with any observable behavior, given appropriate adjustments to background propositional attitudes (Bermúdez, 2009; Morton, 1996, 2003). For example, one may believe that it is raining and desire to stay dry, yet still refuse to open one's umbrella, because one also believes that this action will trigger a bomb. Hence, accurate attribution of propositional attitudes based on observation of finite bouts of behavior, in a timely manner, risks computational intractability, a version of the so-called frame problem in artificial intelligence. In light of these considerations, it is reasonable to hypothesize a more primitive, nonconceptual sociocognitive competence as the basis for most quotidian behavioral prediction.

This nonconceptual sociocognitive competence is most effectively characterized in relational/teleological rather than mentalistic terms (Butterfill & Apperly, 2013; Gergely & Csibra, 2003; Zawidzki, 2011). Natural interpreters predict target behavior by tracking its relations to observable, worldly situations rather than by inferring unobservable, mental causes of behavior. For example, if an interpreter identifies some potentially observable future state as the goal of a target of interpretation, then the interpreter can predict that the target will engage in behavior that constitutes the most efficient means of realizing this state (Gergely & Csibra, 2003). A slightly more sophisticated interpreter might realize that different agents have differential access to information relevant to realizing goal states, where such information access is understood entirely extensionally, in terms of relations to states of affairs the interpreter can observe, for example, lines of sight (Butterfill & Apperly, 2013). Such an interpreter could incorporate sensitivity to such types of information access into her predictions of agent behavior, by relativizing judgments of efficient goal realization to information available to different interpretive targets (Zawidzki, 2011). This still does not require deploying concepts of the mind and mental states: The behavior of targets of interpretation need not be conceptualized as caused by unobservable, subjective mental representations. Because of this, such nonconceptual sociocognitive competencies have the potential to avoid the computational intractability of attributing mental states with holistically mediated connections to behavior. Rather than consider every possible combination of mental states compatible with some agent's situation, such nonconceptual interpreters need focus only on the worldly situations to which interpretive targets can bear a limited number of observable relations—that is, varieties of information access and goal pursuit. Given that in most ecologically plausible contexts, there are a limited number of such behaviorally relevant, observable, teleological, and informational relations between interpretive targets and worldly situations, natural interpreters can thereby avoid the



computational intractability risked in mentalistic attribution and still retain reliable predictive capacities. And it is far more plausible to suppose that nonhuman and human infant interpreters employ such resources than it is to assume that they wield mature concepts of the mind and mental states.<sup>2</sup>

There also seems to be evidence that our mature concepts of mental states more likely serve justificatory functions than predictive ones. For example, Malle, Knobe, and Nelson (2007) investigated a well-known asymmetry in the use of propositional attitude attributions to explain behavior: normal adult subjects are more likely to explain their own than others' behavior in such terms. They tested two hypotheses: (a) we have better epistemic access to our own propositional attitudes; (b) we care more about "impression management" in our own case. When asked to explain behavior of others to whom subjects had varying epistemic access (i.e., friends and family members, directly witnessed strangers, and strangers about whom they were informed by third parties), subjects were *no more likely* to attribute propositional attitudes to those they knew better. However, when asked to make the behavior of even complete strangers look good, subjects were *as likely* to attribute propositional attitudes as when explaining their own behavior. This suggests that our concepts of the propositional attitudes serve more of an impression management or justificatory role than an epistemic role. Such evidence bears out the case that McGeer (1996, 2007, 2015) has been making regarding the justificatory and regulative functions of folk psychology for a number of years. According to McGeer, the many broad disagreements about how folk psychological competence is implemented conceal a deeper consensus: that folk psychology functions primarily to help predict and causally explain behavior. However, McGeer argues persuasively that this is far from evident. There are good reasons to think that folk psychology is efficacious only to the degree that it succeeds in *regulating* behavior to respect justificatory standards and norms.

Thus there is some empirical evidence against one of the basic assumptions that motivate providing a univocal answer to the five questions discussed here. It seems that our mature concepts of the mind and mental states might not play important roles in quotidian behavioral prediction. They might instead be more important to justificatory projects and, hence, as with the concept of a person, provide little insight into the nature of the mind and mental states, considered

---

2. This characterization of the nonconceptual basis for quotidian behavioral prediction and interpretation sounds a lot like IS (Zawidzki, 2012). That is part of my point. IS can be a model of quotidian interpretation, even if quotidian interpretation does not typically involve the deployment of our mature concepts of the mind and mental states. This is especially good news for Dennett if IS is an inadequate analysis of these concepts. More on this follows.



as natural causes of behavior. At this point, the evidence is preliminary, and so I offer these only as possibilities.

#### 4. Weaknesses in IS Analysis of Mature Concepts of Mind

It might seem that the empirical considerations raised in section 3 are not a problem for Dennett's use of IS as a univocal answer to the five questions considered here. After all, my sketch of the nonconceptual basis for our quotidian sociocognitive competence sounds a lot like IS (Zawidzki, 2012). However, part of my point is that IS can be an adequate model of this competence *without also succeeding as an analysis of our mature concepts of the mind and mental states*. This is possible if the latter are not typically used in quotidian social interpretation and prediction. This dissociation between the project of modeling our quotidian social competence and the analysis of our mature concepts of the mind and mental states is key to buttressing Dennett's project against common philosophical complaints, since there are good reasons to doubt his suggestion that IS constitutes an adequate analysis of these concepts. For example, contrary to IS analysis of the concept of belief, it seems the folk assume that there are determinate facts about what people believe beyond what makes best sense out of their actual and counterfactual behavior. Consider Dennett's (1978) example of Sam, the renowned art critic (p. 49). Sam's son is a mediocre artist whose work Sam praises effusively in public. But what does Sam believe? Has his filial love blinded him to the point that he actually believes the mediocre work has value? Or has he become so adept at (self-)deception that his true beliefs remain perfectly concealed until his death? Dennett argues that if there is no behavioral evidence one way or the other, perhaps it is, at bottom, entirely indeterminate what Sam believes. Whether or not Dennett is right about the nature of Sam's mental state, I think it unlikely that he is right about our mature concept of belief. It *seems* that there is a fact of the matter regarding what Sam really believes, whether or not he ever provides, or even could provide, behavioral evidence of it. In any case, this is the intuition that many philosophers attribute to the folk—though appropriate studies would need to be conducted to show that the folk assume such evidence transcending determinacy in belief content. Dennett (1987) himself recognizes this philosophical near-consensus. He argues that some of the most influential contemporary philosophers of mind, including Chisholm, Nagel, Searle, Fodor, Dretske, Kripke, and Burge (p. 295), though they agree on little else, are all committed to what Dennett calls "*intrinsic . . . [or] original intentionality*" (p. 294; original emphasis), the view that mental state contents consist in evidence-transcendent facts.

Dennett's (2014) own distinction between implicit, tacit, manifest image concepts, and the explicit, reflective, "*folk ideology* of the manifest image" (p. 355) is relevant here. It is possible that what we take ourselves to be committed to explicitly and reflectively, regarding the nature of belief, is not what we are actually committed to in practice. As an example, consider folk concepts of freedom of the will. Most philosophers assume that the folk are incompatibilists. If our behavior is entirely causally determined, then it cannot be freely chosen. However, recently, experimental philosophy seems to have thrown some doubt on this assumption (Nahmias & Thompson, 2014). It seems that if questions are posed appropriately and alternative interpretations controlled for, the folk may have compatibilist intuitions after all. Behavior can be determined and freely chosen at the same time. This may be a case in which the explicit, reflective, manifest image concept articulated by philosophers is different from the implicit, tacit, manifest image concept employed by the folk in everyday life. Perhaps something like this is true of folk concepts of propositional attitudes as well. Although philosophers take them to involve commitment to evidence transcending determinacy, in practice, the folk deploy concepts more like those suggested by IS analysis, that is, as fundamentally indeterminate beyond what the behavioral evidence can establish.

However, this strategy for defending IS's analysis of mature concepts of mental states is problematic. It assumes that there are univocal characterizations of implicit, tacit, folk concepts. But there is no reason to suppose that the conceptual resources tacitly wielded by the folk in diverse, practical contexts are systematic or coherent. In fact, there is strong evidence against this. For example, there is evidence that the folk behave as compatibilists in some contexts and incompatibilists in others (Nahmias, 2014, p. 23, n. 10). Similarly, the folk may assume that beliefs have evidence transcending, determinate content in some contexts, and not in others. Which contexts, then, should count as relevant to determining the content of *mature* folk concepts? It seems that the best candidates are those in which we are being most explicit and reflective about our concepts, that is, those in which we are trying to make them as consistent and systematic as possible. In other words, to the extent that it makes sense to talk of univocal, mature folk concepts, our best bet for determining their content is to engage in systematic, philosophical analysis of them. Hence, the philosophical assumption that mature folk concepts are those that result from systematic, philosophical analysis is well motivated. Either we reject the assumption that *there are* mature, univocal folk concepts, or we restrict this notion to concepts that survive rigorous, systematic, philosophical analysis. In either case, Dennett's assumption that IS constitutes an adequate analysis of our mature concepts of the mind and mental states is problematic. It either assumes falsely that there are such univocal, mature concepts, or it gets the best candidates for such concepts wrong.

Furthermore, it is not just the assumption that mental states lack evidence-transcending determinacy that puts IS potentially at odds with our mature concepts of them. There is a host of assumptions that Dennett builds into IS that seem at odds with what the folk assume about the mind and mental states. First, Dennett argues that whether or not an object counts as having a mind or some particular mental state is perspective dependent (1991). Whether or not this is the best version of our concepts of the mind and mental states relative to Dennett's naturalization project, it seems manifestly false that it constitutes an appropriate characterization of our mature concepts of the mind and mental states. Few would agree that whether or not they have beliefs, and which beliefs they have, depend on who is interpreting them and for what purposes. In some places, Dennett argues that folk attribution of desires requires implicit reference to natural selection: organisms are taken to desire states that give them a selective advantage (1987, p. 49). Again, whether or not this is the best version of folk psychology relative to Dennett's naturalization project, it seems to be a non-starter as an analysis of mature folk concepts. There have been millions of human beings who have successfully attributed desires without any awareness of the theory of natural selection, let alone commitment to it. Finally, as has long been appreciated (Stich, 1981), the assumption of practical rationality that Dennett builds into IS seems at odds with folk-psychological practice: the folk routinely attribute rationally incompatible mental states to the same agent. Again, relative to Dennett's naturalization project, it may be appropriate to emphasize the fact that IS enables the tracking of rational patterns of behavior. However, as a characterization of mature, folk-psychological concepts, the claim that a rationality assumption is built into them seems misleading.

In general, the inadequacy of IS as a characterization of our mature concepts of the mind and mental states is a symptom of the very conflation to which I want to draw attention. Because Dennett does not adequately distinguish between the project of characterizing the content of our mature concepts of the mind and mental states and the closely related project of characterizing their natures, especially relative to the goals of naturalizing them and fashioning them into constructs useful to cognitive science, he mistakenly assumes that IS can play all of these roles. Unsurprisingly, therefore, Dennett himself sometimes expresses skepticism about the adequacy of IS as an analysis of our mature concepts of the mind and mental states (1987, p. 55).<sup>3</sup>

---

3. As Amber Ross has pointed out to me, it is doubtful that Dennett's deflationary intentional-stance account of mentalistic terms is intended to capture everything the folk ordinarily mean by "belief," "desire," etc.; instead, he merely wants to capture the aspects of those terms that are worth capturing and saving for a productive cognitive science. If this is true, then perhaps, as I have suggested, this chapter can provoke some final clarification of this issue, given that

## 5. Whither IS?

I believe we should reject the suggestion that IS provides an adequate analysis of our mature concepts of the mind and mental states. The evidence that these concepts play primarily justificatory roles supports this, given Dennett's naturalistic agenda: if IS is supposed to help naturalize the mind, it is probably not a good candidate as an analysis of essentially forensic concepts, which are unlikely to map neatly onto natural kinds. What follows for the other four questions to which Dennett offers IS as an answer?

Is IS a good characterization of what normal humans do when they engage in quotidian behavioral prediction? If the application of our mature concepts of the mind and mental states were central to this competence, then, given the inadequacy of IS as an analysis of these concepts, its role in explaining quotidian behavioral prediction would also be in jeopardy. However, given the evidence that quotidian behavioral prediction relies on considerably less sophisticated capacities, it is possible that IS can help explain such competencies. Even if our sophisticated, reflective concepts of the mind and mental states do not succumb to Dennettian analysis, it is possible that the unreflective capacities for behavioral prediction of infants, nonhuman animals, and normal adults in the heat of seamless, dynamic interaction, do involve something like IS. There is even empirical evidence supporting this conjecture (Gergely & Csibra, 2003; Zawidzki, 2011, 2012).

Gergely and Csibra's "teleological stance" (2003) and Butterfill and Apperly's "minimal theory of mind" (2013) invoked earlier to characterize nonconceptual, quotidian, sociocognitive competence, are readily assimilated to IS. Both deny that most quotidian interpretation requires the attribution of unobservable, internal, concrete, mental causes, focusing instead on observable relations between interpretive targets and worldly states that constitute their goals and information to which they have access. Gergely and Csibra's teleological stance explicitly invokes a notion of instrumental rationality, centered on the efficiency with which behavioral means realize goals. This idea gives determinate content to the assumption of rationality implicit in IS. Furthermore, in defending the hypothesis that the teleological stance is a better characterization of infant sociocognitive

---

Dennett sometimes *seems* to aim for more, in line with the project of his teacher, Ryle. In any case, it is worth clearly differentiating between the projects of adequate analysis and reconceptualization for use in cognitive science, since, as I have pointed out here, many of the most persuasive criticisms of Dennett's project rely on their conflation. Finally, as I will go on to suggest, there are resources within Dennett's overall corpus for providing a richer account of our ordinary mentalistic concepts, that is, by appreciating their regulative uses in the context of what Carol Rovane (1994) calls "the personal stance."

competence than mentalizing alternatives, like the so-called theory-theory and simulation theory, Gergely (2011) points to the fact that infants find agency in an extraordinary variety of objects, many of which bear very little surface resemblance to themselves or other paradigmatic, biological agents, “Young infants are ready to interpret unfamiliar entities such as inanimate objects, abstract 2D figures, humanoid robots, unfamiliar human actions, and even biomechanically impossible hand actions . . . as goal-directed as long as they show evidence of context-sensitive justifiable variation of action obeying the principle of efficiency of goal approach” (p. 87). On the other hand, even a human hand grasping an object “is *not* interpreted as goal-directed . . . if this outcome . . . is achieved as the end result of unnecessary and therefore unjustifiable and inefficient preceding actions, as when a hand first opens a transparent empty box before grasping the target object that is *in front of* the box” (pp. 86–87). Gergely concludes that “what seems to be criterial for attributing intentionality and goal-directedness is evidence indicating the ability for *rational choice* among the accessible action alternatives by reliably performing the most efficient action available to bring about the goal state across changing environmental constraints” (p. 87). It is hard to think of more explicit empirical vindication of IS as a model of, at least infant, quotidian interpretation. From the very start, Dennett has motivated his theory by emphasizing the variety of objects that succumb to IS interpretation, as long as their behavior can be interpreted as instrumentally rational.

Furthermore, the very feature that, as I argued in section 4, makes IS an implausible analysis of our mature concepts of the mind and mental states—that is, its tolerance for indeterminacy of mental state content beyond what the evidence can establish—supports its plausibility as a model of quotidian interpretation.<sup>4</sup> Another way of understanding this point is in terms of the appearance/reality distinction, applied to behavioral phenomena. In proposing IS as both an analysis of our mature concepts of the mind and mental states and as a model of quotidian interpretation, Dennett, in effect, denies that the folk *ever* make a strong appearance/reality distinction, when it comes to interpreting behavior. Of course, he does not deny that folk interpreters think of their interpretations as fallible. The point, rather, is that, if IS’s characterization of folk interpretive competence is correct, then the notion that there is a mental reality that may vary independently of all possible behavioral evidence should make no sense to

---

4. This seems paradoxical, but only if one assumes that our mature concepts of the mind and mental states play important roles in most quotidian interpretation. Since there is good reason to deny this, it is possible that most quotidian interpretation does *not* presume evidence-transcending determinacy in mental state content, *even if our mature concepts of the mind and mental states do*.

the folk. On Dennett's view, once all the behavioral evidence is in and supports one IS interpretation, or more, over others, there is no further fact of the matter about what the target of interpretation really thinks. This is precisely where most philosophical analyses of our mature concepts of the mind and mental states disagree with Dennett's analysis. It seems obvious that *when it comes to our mature concepts of the mind and mental states*, we do make a strong distinction between behavioral appearance and mental reality: it is possible, relative to these concepts, for there to be determinate facts about an agent's mental states that completely transcend all possible behavioral evidence. However, it is precisely the denial of this that makes IS a plausible model of *the actual cognitive resources we use in most quotidian behavioral prediction*. For, if our task is just behavioral prediction, of what possible use could such a strong behavioral appearance/mental reality distinction be? Why waste time trying to get mental attributions right, beyond what works for predictive purposes? If two interpretations are equally compatible with all possible behavioral data, and make exactly the same behavioral predictions, distinguishing between them should be of no further interest, if quotidian behavioral prediction is our only goal. Thus, the very feature of IS that makes it implausible as an analysis of our mature concepts of the mind and mental states—its eschewal of a strong behavioral appearance/mental reality distinction—actually makes it more plausible as a model of quotidian interpretation.

There are thus both empirical and conceptual reasons for retaining IS as a model of quotidian interpretation, even if it fails as an analysis of our mature concepts of the mind and mental states. What of the other three questions? Can IS survive as an account of the nature of mental states, as a key to the naturalization of the mental, and as an important part of the methodology of cognitive science, even if it fails as an analysis of our mature mentalistic concepts? These four questions are related, and the role of IS in answering them depends on how one understands their interrelations. If one sees the nature of mental states as exhaustively determined by our mature concepts of them, then the failure of IS to adequately analyze the latter would seem to disqualify it as an account of the former. Some concepts are like this. We do not think there is anything more to the nature of fictional entities like, for example, hobbits than what our concepts of them stipulate. But other concepts are not like this. Our ordinary concepts were very poor guides to the nature of, for example, heat, water, or lightning. Thus, a lot depends on whether or not we think of the mind and mental states as natural kinds.<sup>5</sup> If we do, then there is *no* reason to think that our

---

5. I mean something very minimal by "natural kinds" here: merely the possibility that a folk or scientific concept can succeed in picking out a real object or category despite encoding assumptions that are largely wrong about it.

mature concepts of the mind and mental states put strong constraints on the nature of the mind and mental states. In that case, even if it fails as an analysis of our mature concepts of them, IS may play a role in a correct characterization of the nature of the mind and mental states. It all depends on whether or not it plays an important role in the project of naturalizing the mind, and this, arguably, depends on whether or not it plays an important role in the scientific study of cognitive phenomena.

Once the logical geography is rendered in this way, I think that Dennett's project retains a lot of its promise. Even if IS is inadequate as an analysis of our mature concepts of the mind and mental states, there are good reasons to think that it gets something importantly right about the "System 1" sociocognitive capacities that underlie most successful quotidian behavioral prediction. The interpretive and predictive feats of nonhuman animals, human infants, and human adults in the heat of seamless, dynamic interaction are effectively explained in terms of IS: our sensitivity to the goals, information access, and instrumental rationality of agents with which we interact effectively explains our capacities to predict them. The success of such quotidian interpretive feats is best made sense of if we assume that IS gets something right about the nature of agency, that it tracks a "real pattern," in Dennett's (1991) sense. Even if not all such real patterns involve natural, or scientific kinds, for example, the "geocentric stance" tracks only "parochial" patterns in the motion of celestial bodies that happen to be accessible to the earthbound, Dennett has provided a good reason for thinking that the real patterns tracked via IS are important "joints" in nature: natural selection will tend to produce systems that pursue their goals efficiently in the light of relevant information. So, unlike, say, the "geocentric stance," IS will track real patterns wherever natural selection has been at work. Since natural selection is the only known natural process capable of producing intelligent pattern trackers, the patterns visible via IS will be available to any intelligent pattern trackers. Because of their ubiquity and importance, the real patterns produced by natural selection, and tracked in quotidian life using IS are phenomena worthy of scientific investigation, particularly by the biological, cognitive, and social sciences. Since IS is an effective means of tracking such patterns, some version of it will likely prove very useful to the methodologies of these sciences, as Dennett has argued.

## 6. Conclusion

It is unlikely that there is a correct, univocal response to the following five explanatory challenges in philosophy of mind and cognitive science: (a) the analysis of our mature concepts of the mind and mental states; (b) the correct theory of our



quotidian competence at predicting each other's behavior; (c) the correct theory of the nature of the mind and mental states; (d) the appropriate strategy for naturalizing the mind and mental states; (e) the correct methodology for cognitive science. In some places, Dennett defends IS as a response to all five of these challenges, but it fails in this for two main reasons.

First, there is an entirely general reason why any univocal response to these five challenges will likely fail. There is increasing evidence that our mature concepts of the mind and mental states play little role in successful quotidian behavioral prediction, and are better conceptualized as playing forensic roles. This is highly problematic given certain plausible and widespread assumptions. If our mature concepts of the mind and mental states do not play important roles in successful quotidian behavioral prediction, then there is little reason to suppose that they are accurate depictions of the nature of the mind and mental states. Furthermore, if they are primarily forensic constructs, there is no more reason to suppose that they are naturalizable, and important to cognitive science, than there is to suppose that our mature concept of a person is.

The second reason why Dennett's univocalizing strategy fails is that IS is an inadequate analysis of our mature concepts of the mind and mental states. These concepts appear to assume evidence transcending determinacy of mental state content (i.e., a strong appearance-reality distinction applied to behavioral phenomena), perspective-independent facts about the instantiation of mental states, no commitment to the theory of natural selection, and no necessary commitment to the practical rationality of objects of mental state attribution, all of which contradict IS analysis of mentalistic concepts. Does this imply that Dennett has absolutely nothing interesting to contribute to the project of explicating our mature concepts of the mind and mental states? Not necessarily. Besides IS, Dennett has occasionally discussed an interpretive framework that Carol Rovane has termed the "personal stance" (1994). This is the stance that we take toward persons—that is, members of our ethical community. It is therefore plausible that this stance plays the kinds of regulative roles that McGeer (1996, 2007, 2015) has emphasized. Perhaps there are good reasons, relative to this regulative function, that we take the mental states of *persons* to have contents with evidence-transcending determinacy, to constitute perspective-independent facts, and to be independent of practical rationality or the truth of natural selection. On this view, our mature concepts of the mind and mental states belong to the personal stance, not to IS. Because their role is primarily regulative, they are of little relevance to understanding the nature of the mind, considered as a natural object studied by the cognitive sciences.

Ironically, it is precisely the failure of IS as an analysis of mature mentalistic concepts that may rescue its plausibility in playing the other roles Dennett



envisions for it. If mature mentalistic concepts do not play an important role in successful, quotidian behavioral prediction, then the failure of IS's analysis of the former does not preclude its use as a model of the latter. And, indeed, there are good empirical and conceptual reasons for treating IS as a plausible model of our quotidian interpretive competence, despite, and perhaps even because of its failure as an analysis of mature mentalistic concepts. If IS does prove to be an accurate model of successful behavioral prediction, it then becomes plausible to assert that it gets something right about the nature of minded agents. Given a commitment to naturalism, this would not be possible unless it also had important roles to play in both naturalization projects and cognitive science. Indeed, Dennett's own explanation of why IS works, that is, in virtue of tracking real patterns generated by natural selection, points the way to explaining the important roles it can play in these latter two endeavors.

## Acknowledgments

I wish to thank Bryce Huebner for organizing this wonderful volume and giving me the opportunity to contribute, and for excellent, detailed, incisive comments on an earlier draft that undoubtedly helped improve the final product. Any remaining weaknesses are entirely my responsibility.

## Works Cited

- Akins, K. (1996). Of sensory systems and the "aboutness" of mental states. *Journal of Philosophy*, 93, 337–372.
- Apperly, I. A. (2011). *Mindreaders*. New York, NY: Psychology Press.
- Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, 116, 953–970.
- Bermúdez, J. L. (2009). Mindreading in the animal kingdom. In R. Lurz (Ed.), *Animal minds* (pp. 145–164). Cambridge, UK: Cambridge University Press.
- Bugnyar, T., Reber, S. A., & Buckner, C. (2016). Ravens attribute visual access to unseen competitors. *Nature Communications*, 7. Retrieved from doi:10.1038/ncomms10506
- Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind and Language*, 28, 606–637.
- Clayton, N. S., Dally, J. M., & Emery, N. J. (2007). Social cognition by food-caching corvids: The western scrub-jay as a natural psychologist. *Philosophical Transactions of the Royal Society B*, 362, 507–552.
- Dennett, D. C. (1978). *Brainstorms*. Cambridge, MA: MIT Press.
- Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

- Dennett, D. C. (1991). Real patterns. *Journal of Philosophy*, 88, 27–51.
- Dennett, D. C. (1994). Self-portrait. In S. Guttenplan (Ed.), *A companion to the philosophy of mind* (pp. 236–244). Oxford, UK: Blackwell.
- Dennett, D. C. (2009). Intentional systems theory. In B. McLaughlin, A. Beckermann, & S. Walter (Eds.), *The Oxford handbook of the philosophy of mind* (pp. 339–350). Oxford, UK: Oxford University Press.
- Dennett, D. C. (2014). *Intuition pumps and other tools for thinking*. London: Penguin.
- Fodor, J. A. (1987). *Psychosemantics*. Cambridge, MA: MIT Press.
- Gergely, G. (2011). Kinds of agents: The origins of understanding instrumental and communicative agency. In U. Goshwami (Ed.), *Blackwell handbook of childhood cognitive development* (2nd ed., pp. 76–105). Oxford, UK: Blackwell.
- Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naive theory of rational action. *Trends in Cognitive Sciences*, 7, 278–292.
- Hare, B., Call, J., & Tomasello, M. (2006). Chimpanzees deceive a human competitor by hiding. *Cognition*, 101, 495–514.
- Hutto, D. (2008). *Folk psychological narratives*. Cambridge, MA: MIT Press.
- Kovacs, A. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others-beliefs in human infants and adults. *Science*, 330, 1830–1834.
- Malle B. F., Knobe J., & Nelson, S. E. (2007). Actor-observer asymmetries in behavior explanations: New answers to an old question. *Journal of Personality and Social Psychology*, 93, 491–514.
- Marr, D. (1982). *Vision*. New York, NY: Freeman.
- McGeer, V. (1996). Is “self-knowledge” an empirical problem? Renegotiating the space of philosophical explanation. *Journal of Philosophy*, 93, 483–515.
- McGeer, V. (2007). The regulative dimension of folk psychology. In D. D. Hutto & M. Ratcliffe (Eds.), *Folk psychology re-assessed* (pp. 137–156). Dordrecht, The Netherlands: Springer.
- McGeer, V. (2015). Mind-making practices: The social infrastructure of self-knowing agency and responsibility. *Philosophical Explorations*, 18, 259–281.
- Morton, A. (1996). Folk psychology is not a predictive device. *Mind*, 105(417), 119–137.
- Morton, A. (2003). *The importance of being understood*. London: Routledge.
- Nahmias, E. (2014). Is free will an illusion? Confronting challenges from the modern mind sciences. In W. Sinnott-Armstrong (Ed.), *Moral psychology* (Vol. 4, pp. 1–25). Cambridge, MA: MIT Press.
- Nahmias, E., & Thompson, M. (2014). A naturalistic vision of free will. In E. Machery & E. O’Neill (Eds.), *Current controversies in experimental philosophy* (pp. 86–103). New York, NY: Routledge.
- Rovane, C. (1994). The personal stance. *Philosophical Topics*, 22, 351–396.
- Shoemaker, D. (2014). Personal identity and ethics. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring Ed.). Retrieved from <http://plato.stanford.edu/archives/spr2014/entries/identity-ethics/>

- Stich, S. (1981). Dennett on intentional systems. *Philosophical Topics*, 12, 38–62.
- Zawidzki, T. W. (2011). How to interpret infant sociocognitive competence. *Review of Philosophy and Psychology*, 2, 483–497.
- Zawidzki, T. W. (2012). Unlikely allies: Embodied social cognition and the intentional stance. *Phenomenology and the Cognitive Sciences*, 11, 487–506.

## 2.2 REFLECTIONS ON TADEUSZ ZAWIDZKI

Daniel C. Dennett

Tad Zawidzki has done a fine review of my writing about the intentional stance (IS), and uncovered no less than five separable roles for it. With the hindsight he provides, I can agree down the line with all the main points of his analysis: there are indeed (at least) five claims that I have made about the multifarious roles of the IS, and they are all separable in conception if not easily in practice. Articulating these different possibilities clearly is justification enough for this paper, but in fact it goes further, opening up space for several themes that have not been given the attention they are due.

Zawidzki is right that I have ventured all five proposals and vacillated on which might be the core or foundation of the IS. And he is right that I have seen all five as fitting together so well there was no need to rank order them from obligatory to optional. And now that he raises the issue so forcefully, I see that I also agree with him that I might well drop the first role; that is, I might jettison the IS as an analysis of “our mature concepts.” (It could be the ladder I throw away once I have climbed it.) I agree, that is, that “IS can be an adequate model of [our quotidian competence] *without also succeeding as an analysis of our mature concepts of the mind and mental states*” (p. 46). So much the worse, then, for our “mature concepts,” which, on this reading, emerge as not just mature but obsolete, relics of an earlier *Zeitgeist* kept in circulation by energetic discussion and intuition-mongering by philosophers. As Zawidzki says, “It *seems* that there is a fact of the matter about what Sam really believes, whether or not he ever provides, or even could provide, behavioral evidence of it.” So it does *seem*, and this expresses what Quine calls the “museum myth” of meaning, the idea that there *must be* some internal physical structure or linkup or registration that settles the question of what somebody believes, or means—if only we knew how to look for it.

Quine's (1960) famous—or I guess I should say infamous—demolition of the museum myth is the central thought experiment of *Word and Object*, demonstrating the “indeterminacy of radical translation.” I say “demonstrating” advisedly, knowing that I am in the minority in pledging allegiance to this quintessentially Quinian doctrine and wanting to shock the complacent majority into reconsidering their lightly considered verdict, especially the younger generation of philosophers of language and mind who “learned” from their professors that this was Quine's bridge too far, an extreme view that they didn't have to take seriously. It has always baffled and frustrated me that so few of my contemporaries could get their heads around the indeterminacy of radical translation, and many, it seemed to me, went on to teach courses that conveyed a hobbled, nay eviscerated, Quine to their students.

Another excellent recent essay on the IS, by Richard Dub (2015), joins Zawidzki in neglecting the role I take the IS to play in explaining my allegiance to Quine's indeterminacy principle. (See Dennett, 2015, for my response to Dub.) Alerted by Dub's paper, I have come up with a hunch. I suspect that I have been assuming something “goes without saying” that manifestly needs saying. I have overestimated the extent of Quine's influence on the field. If this is what I have done, I do have some shred of an excuse: with superstars Donald Davidson and David Lewis firmly on my side, I could be forgiven, perhaps, for thinking that I need not rehearse the obvious, when they, and our shared mentor, had said it all so well, so authoritatively.<sup>1</sup>

So I am tempted to surmise that Zawidzki graciously ignored my Quinean excesses, a misapplication of the principle of charity since I most definitely want to add a sixth role to his five, and it is arguably the most important role of all: the IS, with its rationality constraint, explains why in principle—if seldom if ever in fact—there is indeterminacy of radical translation/interpretation: there can always be a tie for first between two competing assignments of meaning to the behavior (and behavior-controlling internal states) of an agent, and no other evidence counts.<sup>2</sup> Those who analyze our “mature concept” of belief, for instance,

---

1. Donald Davidson had been out of town when I presented “Intentional Systems” at Princeton in the fall of 1970, but in the summer of 1971, at the famous UC Irvine summer workshop on mind and language organized by Davidson and Harman, Davidson told me that his Princeton students kept telling him that what he was saying was strongly reminiscent of what I had been saying in *Content and Consciousness*, but that he hadn't got around to reading the book yet. When Davidson presented “Belief and the Basis of Meaning” and David Lewis presented “Radical Interpretation” at the 1973 UConn conference on language, intentionality, and translation theory (see *Synthese*, 1974, special issue), there was some discussion of the similarities between his view, Davidson's, and mine in “Intentional Systems,” and Quine was there to offer his benediction to us all. (See also Quine's, 1974, wonderful rebuttal of Michael Dummett's (1974) carefully argued canvassing of the possibilities at the UConn conference. Scholars may regret that the three of us never got around to exploring our agreements and residual disagreements, but we all had better things to do at the time.

2. Isn't this just *behaviorism*? Yes, of a sort. The good sort that all naturalists need to espouse, in the end. Suppose I'm wrong. Then let there be an internal Language of Thought, as direct

may discover that there is a presumption built into it that there is always a matter of fact about just which proposition is believed, but we may ask if there isn't a *better* concept of something very like belief, playing much the same role in explanation (and prediction, and action guidance), which eschews this demand. As in so many philosophical domains, there is no bright line separating analysis from revision. Or at least, so I have often claimed: should one say that consciousness exists, but isn't what people (mature people) think it is, or that consciousness is an illusion? Ditto re free will. Should we say that atoms don't exist, or that they do but they aren't what people think they are (indivisible, by definition)? The trouble with "mature" concepts is that they tend to fossilize old science, which may be obsolete.

It seems clear to me that the very practice of philosophical analysis illustrates this central feature of the IS: as an interpretive strategy, it aims to find the best, most defensible version of whatever the topic is, and wherever there are two competing interpretations with no clear advantage to one of them, we shouldn't make the mistake of insisting that there has to be a fact of the matter about which interpretation is "the truth about" the topic. Sometimes, in the end, one interpretation is revealed to be substantially better, all things considered. But don't count on it. So Amber Ross (see Zawidzki's note 3) is right: I want to capture the aspects of the folk concepts "worth capturing and saving for a productive cognitive science." Anybody who wants to insist that what I end up with doesn't capture everything the folk ordinarily mean will find me agreeing wholeheartedly. I have to launder out the ideology, the dualism, the essentialism, the Cartesianism, the museum myth of meaning, before proceeding. And I don't want to stop at "productive cognitive science"; I see an enlarged role for the IS in evolutionary biology, in accounting for the "free-floating rationales" that inform the functional analyses of organs and behaviors and other features of the living world. That leaves me a further burden: explaining how the folk can be so successful while using concepts that are so laden with indefensible ideology, a project undertaken in *BBB*, along with the explanation of how and why the IS evolved in the first place. Zawidzki sees the centrality of natural selection to my account of the IS and nicely articulates the outline of my account: "natural selection will tend to produce systems that pursue their goals efficiently in the light of relevant information."

---

a touchstone of the *real* meaning of an assertion, or the *real* content of a belief, as one could pray for. "See!" you say, "written right here in the brain, in English: My son is a brilliant artist." What makes you so sure that it's English that you are reading? The proof of *that* assumption would take you back out to the interpretation of behavior (what else, God whispering in your ear?), and if there can be two competing interpretations of the external behavior, there can be two competing interpretations of the internal behavior.

What about the “forensic role” of the “personal stance” (Dennett, 1976; Rovane, 1998)? Zawidzki sees that if the concepts of belief, intention, desire, etc., that we wield in our legal and moral practices are the “mature concepts” that I have either discarded or laundered all out of shape, there is a serious tension between the IS and our conception of ourselves as moral agents. Yes, and this is what generates the current debate about whether or not “neuroscience” shows that “free will is an illusion.” It is undeniable that neuroscience shows either that free will doesn’t exist or that it does but isn’t what people think it is. That is to say, I am willing to concede for the sake of argument that the “mature concept” of free will is some mishmash of libertarian agent-causation or other indeterministic fantasy, but then people are wrong to think that they need that metaphysical baggage to sustain a valuable concept of responsibility or desert. The IS has sufficient resources to distinguish morally competent agents, in terms of their dispositions to believe (register) and desire (prefer), the extent of their knowledge and powers of imagination, their ability to be “moved by reasons.” Why don’t more philosophers see this? Because, I suspect, they start their project of analysis of the mature concepts without any scientific axes to grind, and hence are not motivated to put any strain on the ordinary concepts, to see if they can be stretched or weakened or otherwise adjusted so as to fit well with, say, neuroscience and evolutionary biology. They take the terms at face value, in short, and see what kind of theory or model of a responsible agent they can construct without any procrustean pinching of the mature concepts. That creates a philosophico-anthropological artifact: The Everyday Concept of a Moral Agent with Free Will. That’s a valuable thing to have as a starting point, or a reference point, or a beacon, but it doesn’t have to be taken as the last word on what the concepts are, or what the phenomena are that the concepts purport to account for.

## Works Cited

- Davidson, D. (1974). Belief and the basis of meaning. *Synthese*, 27, 309–323.
- Dennett, D. (1976). Conditions of personhood. In A. O. Rorty (Ed.), *The identities of persons* (pp. 175–196). Berkeley: University of California Press.
- Dennett, D. C. (2015). Not just a fine trip down memory lane: Comments on the essays on content and consciousness. In C. Muñoz-Suárez & F. De Brigard (Eds.), *Content and consciousness revisited* (pp. 199–220). New York, NY: Springer.
- Dub, R. (2015). The rationality assumption. In C. Muñoz-Suárez & F. De Brigard (Eds.), *Content and consciousness revisited* (pp. 93–110). New York, NY: Springer.
- Dummett, M. (1974). The significance of Quine’s indeterminacy thesis. *Synthese*, 27, 351–397.

- Lewis, D. (1974). Radical interpretation. *Synthese*, 27, 331–344.
- Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.
- Quine, W. V. O. (1974). Comment on Michael Dummett. *Synthese*, 27, 399.
- Rovane, Carol. (1998). *The bounds of agency: An essay in revisionary metaphysics*. Princeton, NJ: Princeton University Press.



## 3.1 MEMORY AND THE INTENTIONAL STANCE

Felipe De Brigard

There are many topics one can't help but associate with Daniel Dennett: consciousness, free-will, evolution, intentionality, religion. But in discussions of memory, his name may not come up as readily. He mentions the role of memory in consciousness (Dennett, 1978, 1991) and dreaming (1976), for instance, but only one paper addresses memory directly, *Mining the Past to Construct the Future: Memory and Belief as Forms of Knowledge*, which he coauthored with his former postdoc, psychologist Chris Westbury (Westbury & Dennett, 2000), and published in a relatively obscure volume on memory and belief. This paper has been cited just 37 times, with fewer than 10 citations in philosophy venues. This is unfortunate, in my opinion, because Westbury and Dennett (henceforth W&D) delineate a viable and coherent view of episodic memory that has received substantial support during the last decade and a half of scientific research. Or so I argue here.

I begin by recapitulating W&D's article in order to highlight their three key theses: a *functional* thesis about how memory works, a *computational* thesis about the process of remembering, and a *metaphysical* thesis about the ontological status of memories. In the rest of the chapter I argue that these three theses not only are consistent with one another, but also constitute a coherent view on episodic memory and remembering that has received strong support from different research areas. In section 2, I review recent evidence from cognitive psychology, neuroscience, and neuropsychology that lends credence to W&D's functional thesis. In section 3, I review recent work in computational psychology and cognitive science that shows how W&D's computational thesis can be modeled, and that their way of thinking about memory's computational underpinnings may be a good fit for extant data. Finally, in section 4, I offer a reading of W&D's metaphysical thesis that helps to pull together this evidence into a coherent and explanatorily powerful view of the nature of episodic memory, a view according to which remembering is best understood from the intentional stance.

## 1. An Opinionated Reconstruction of Westbury and Dennett (2000)

The editors of the volume in which W&D's article appears characterize their contribution as "a useful historical overview of the ways in which philosophers have attempted to delineate the boundary conditions of both memory and belief" (Schacter & Scarry, 2000, p. 4). The authors, though, see their aim as that of clarifying "some conceptual confusions that spring from the way in which we use [the terms 'memory' and 'belief'] in informal discourse" (Westbury & Dennett, 2000, p. 12). But I think their article offers more than this. My objective, however, isn't to summarize their paper, but rather to reconstruct it as I read it—essentially, as an exercise in reverse-engineering memory (Dennett, 1994).

W&D's chapter starts off with the observation that while some events leave long-term traces, others don't. Events that do not leave long-term traces are "inert historical facts": their past occurrence makes no difference now. For a past event to make a difference now it needs to leave a long-term trace that has the potentiality of becoming operational when it is needed. In the case of experienced events, W&D tell us, "Memory in the fundamental sense is the ability to store useful information and to retrieve it in precisely those circumstances and that form which allow it to be useful" (2000, p. 13). There is a strong forward-looking flavor to this understanding of memory. For the recording of traces is not haphazard. Somehow memory must anticipate what is most likely to become useful at a later time. It would be strategically damaging to save facts that are unlikely to make a difference, as that would be a waste of valuable cognitive storage. Memory must be attuned to store not merely what happens, but also what is likely to be needed in the future.

The hypothesis that our minds are essentially anticipatory devices has an eminent tradition, which, according to one recent account (Hohwy, 2013), includes Ibn al Haytham (ca. 1030; Sabra, 1989), Kant (1781), and, more explicitly, von Helmholtz (1866). More recently, many authors have bolstered such a view of the mind by supporting the related hypothesis that brains evolved to essentially become anticipatory machines (Clark, 2016; Llinás, 2001), a hypothesis that W&D seem to wholeheartedly embrace:

The whole point of brains, of nervous systems and sense organs, is to produce future, to permit organisms to develop, in real time, anticipations of what is likely to happen next, the better to deal with it. The only way—the only nonmagical way—organisms can do this is by prospecting and mining the present for the precious ore of historical facts, the raw materials that are then refined into anticipations of the future. (2000, p. 12)

This general principle, W&D remind us, is especially true of the mechanistic operations of our memory system. Memory should not be seen merely as a passive receptor of information about the past but as an active producer of anticipatory information about the future (see Ingvar, 1985, for a related idea). Moreover, these anticipations of the future—W&D conjecture—are constructed out of the same material as our memories of the past. Together, these two considerations amount to what I take to be W&D's functional thesis:

*Functional thesis: Our memory system not only processes information about past events but also uses this information to construct useful anticipations of possible future events.*<sup>1</sup>

Thinking of memory in these terms leads to an inevitable question: How can such a system know in advance which information will be useful in the future? W&D's suggestive response relates to another eminent view, usually associated with Bartlett, according to which the encoding and the retrieval of a particular event are influenced by previously acquired knowledge. More precisely, Bartlett's idea was that memory involves "an active organization of past reactions" (1932, p. 201), so that every new experience is encoded and retrieved not only as an individual event, but also as related to knowledge of similar events we have experienced and encoded. These *schemas* relieve memory from encoding every detail of each event as a unique occurrence—a feat that, given perceptual and attentional limitations, is simply impossible. In turn, when the time comes to retrieve information, nonencoded details are easily reconstructed in accordance with both schematic knowledge and knowledge of the present situation; or, as W&D

---

1. Further textual support for the functional thesis can be found in one of the very few passages in which Dennett talks about episodic memory in *Consciousness Explained* (1991), where he suggests that episodic recollection may have coevolved with our capacity for anticipation:

These techniques of representing things to ourselves permit us to be self-governors or executives in ways no other creature approaches. We can work out policies way in advance, thanks to our capacity for hypothetical thinking and scenario-spinning; we can stiffen our own resolve to engage in unpleasant or long-term projects by habits of self-reminding, and by rehearsing the expected benefits and costs of the policies we have adopted. And even more important, this practice of rehearsal creates a memory of the route by which we have arrived at where we are (what psychologists call *episodic memory*), so that we can explain to ourselves, when we find ourselves painted in the corner, just what errors we made. . . . The developments of these strategies permitted our ancestors to look farther into the future, and part of what gave them this enhanced power of anticipation was an enhanced power of recollection—being able to look farther back at their own recent operations to *see* where they made their mistakes." (p. 278)

put it, “What we recall is not what we actually experienced, but rather a reconstruction of what we experienced that is consistent with our current goals and our knowledge of the world” (2000, p. 19).

This reconstructive view of memory implies that memories do not lie dormant and immutable after encoding, but are reconstructed, at the time of retrieval, in accordance with the constraints dictated by previously acquired schemas as well as present conditions of recall:

The apparently stable objects of memory—the representations of the things being recalled—are not retrieved from some Store-house of Ideas where they have been waiting intact, but rather are constructed on the fly by a computational process. As H. R. Maturana (1970) wrote . . . “memory as a storage of representations of the environment to be used on different occasions in recall does not exist as a neurophysiological function” (p. 37). What we call recollection can never be more than the most plausible story we come up with (or, perhaps, only a story which is plausible enough) within the context of the constraints imposed by biology and history. (Westbury & Dennett, 2000, p. 19)

This last point—that recollection involves the reconstruction of “the most plausible story we come up with”—suggests that the computational process of reconstructive retrieval W&D have in mind operates probabilistically. Moreover, it suggests that the plausibility of the reconstruction is somehow tied to biology as well as individual history:

The assessment of what constitutes an acceptable reconstruction of the past must be dynamically computed by an organism under the constraints imposed by its built-in biological biases and the history of the interaction of those biases with the environment in which the organism has lived. (p. 19)

If we take these historical interactions as referring, at the very least, to the recording of statistical regularities of past events during the generation of the aforementioned knowledge structures or schemas, then we get close to what I take to be W&D’s computational thesis:

*Computational thesis: Our memory system is sensitive to the statistical regularities of the information it encodes, which along with the individual’s limitations and current goals, constrain retrieval in a probabilistic manner.*

At this point, W&D change gears and talk about how *philosophers of mind* typically talk about belief, which differs from the way ordinary folk use the term “belief” (p. 20). According to W&D, when philosophers of mind talk about belief, they often mean some kind of sentence-like informational structure in the brain. However, since such a view is unlikely (Dennett, 1978, 1987; Westbury & Dennett, 2000, pp. 20–24), W&D propose to understand belief from the intentional stance:

Instead of continuing the attempt to define a belief as an entity that an organism might have or not have (in the concrete, binary-valued way that a library can have a particular book or a poem can have a particular line), a belief must be defined in terms of the circumstance under which a belief could be justifiably *attributed* to an organism. What is meant when it is asserted that an organism has a belief, we propose, is that its behavior can be reliably predicted by ascribing that belief to it—an act of ascription we call taking the intentional stance. (p. 24)

This perspective has two immediate consequences. First, what makes an attribution of belief true has nothing to do with whether or not there is a particular brain structure encoding the content expressed by the belief. And second, two individuals—or the same individual at two times—can have the same belief even if what goes on in their brains at each time is different (cf., De Brigard, 2015).

W&D go on to suggest that these considerations also apply to memories. One can say, of a certain individual, that she’s remembering that *p* even if there is no structure in the brain exclusively dedicated to encoding the content *p*. Moreover, seeing memories from the intentional stance allows us to say, of two different individuals, that they are remembering the same memory even if what goes on in their brains is different. Finally, it allows us to say that one can remember the same memory today that one remembered yesterday, even if the neural processes that are engaged at each time are different. Adopting the intentional stance toward memories, therefore, invites us to think of them from the point of view of what I take to be W&D’s metaphysical thesis:

*Metaphysical thesis: memories do not exist as subpersonal-level brain structures encoding particular intentional contents, but rather as personal-level psychological phenomena only accessible from the intentional stance.*

Perhaps my reconstruction of W&D’s article is mostly wishful reading, but I see them as offering more than “conceptual clarifications” of “memory” and “belief.” I take them to be making substantive claims about memory, given its

role in anticipation, reconstructive character, and personal rather than subpersonal nature. The beauty of this reverse-engineering approach is that the resultant theses are now, 15 years later, poised for empirical scrutiny, and the verdict of the tribunal of experience favors W&D's case.

## 2. The Functional Thesis

According to the functional thesis, memory must both process information about past events and employ this information to construct useful anticipations of possible future events. This general hypothesis had received scattered empirical support when W&D wrote about it. Korsakoff (1889/1996) had described an amnesic patient who was both incapable of remembering past experiences and unable to imagine new plans. And Talland (1965) had described instances of chronic Korsakoff's syndrome when patients had difficulty imagining possible future events and remembering past personal experiences. This also appears to have been the case with the amnesic patients H. M. (Buckner, 2010), and K. C. In an oft-quoted passage, Tulving (called "E. T."; 1985, p. 4) asks K. C. (called here "N. N.") what he will be doing the following day:

- E. T: "Let's try the question again about the future. What will you be doing tomorrow?" (There is a 15-second pause.)  
 N. N: smiles faintly, then says, "I don't know."  
 E. T: "Do you remember the question?"  
 N. N: "About what I'll be doing tomorrow?"  
 E. T: "Yes. How would you describe your state of mind when you try to think about it?" (a 5-second pause)  
 N. N: "Blank, I guess."

K. C.'s inability to remember past personal events and to imagine possible future episodes inspired Tulving to argue that our capacities for episodic memory and for projecting ourselves into the future were underwritten by a general capacity for "mental time travel."

Stronger support emerged in the early 2000s, as a number of studies suggested that episodic memory and future thinking share common neural substrates. One of the strongest pieces of neuropsychological evidence came from patient D. B., who displayed a profound deficit in episodic memory but normal performance in other cognitive tasks, including semantic memory (Klein, Loftus, & Kihlstrom, 2002). When asked to imagine possible personal future events, D. B. drew a blank, leading Klein and colleagues to posit a common mechanism for mental time travel. Around that same time, Atance and O'Neill (2001) reported evidence

of common developmental trajectories for episodic memory and future thinking. And Okuda and colleagues (2003) presented Positron Emission Tomography (PET) data showing common engagement of medial temporal lobe (MTL) structures during episodic memory and future thinking, in contrast with a control task requiring semantic retrieval. Behavioral studies revealed further parallels between episodic memory and future thinking. For instance, D'Argembeau and Van der Linden (2004, 2006) showed similar effects of valence, temporal distance, and emotion regulation for episodic memory and future thinking. Similarly, Spreng and Levine (2006) showed that the temporal distribution of episodic past and future thoughts could be modeled on a logarithmic scale because it was more common for subjects to generate events closer to the moment of the simulation than more remote events, suggesting common effects of temporal distance for both episodic memory and future thinking.

Finally, in 2007, three papers provided more conclusive evidence in favor of the claim that some common neural mechanisms are required to remember our past and to imagine possible personal future events. In the first, Szpunar, Watson, and McDermott (2007) asked subjects to remember a past personal event, imagine a possible personal future, or imagine a fictitious event of no relevance for them while undergoing functional magnetic resonance imaging (fMRI). Their analysis revealed that thinking about a personal past event and a personal future event, but not thinking about a nonpersonal event, commonly engaged the medial prefrontal cortex (mPFC)—mainly BA10—the posterior cingulate cortex (pCC), MTL, and cuneus. In the second study, also employing fMRI, Addis, Wong, and Schacter (2007) cued participants to either remember a past personal episode, imagine a possible personal future event, or create a sentence. Consistent with previous results, they found remarkable overlap between episodic memory and future thinking, as compared to sentence creation, in regions known to be associated with episodic recollection (Cabeza & St. Jacques, 2007; St. Jacques & De Brigard, 2015). Notably, this common engagement occurred in many of the same regions found by Szpunar et al. (2007), with the only difference being the finding of greater hippocampal activity during episodic memory and future thinking relative to the control condition, suggesting both that the hippocampus could have been involved in the nonpersonal simulations in Szpunar et al.'s study, and that the hippocampus might be indispensable for mental time travel.

The observation that the hippocampus is critical for both episodic memory and future thinking received even stronger support from the third study, where Hassabis and Maguire (2007) asked five patients with hippocampal amnesia, and ten healthy controls, to imagine possible new experiences in response to verbal cues (e.g., "Imagine you're lying on a white sandy beach in a beautiful tropical bay"). Participants' descriptions of their simulations were transcribed and



coded to assess how rich, detailed, and spatially coherent they were. The results of this study clearly showed that amnesic patients' descriptions of their simulations were less rich, contained fewer details, and were less spatially coherent than the descriptions produced by controls. These results were further replicated on different patients by Race, Keane, & Verfaellie (2011), with the additional demonstration that the deficits in the descriptions were independent of the patient's narrative abilities.

Related evidence regarding the role of the hippocampus in mental time travel comes from the animal literature. Since the finding of hippocampal pyramidal cells whose receptive fields are sensitive to spatial locations, researchers have hypothesized that such "place cells" generate the internal maps we rely on for navigation (O'Keefe & Nadel, 1978). These place cells also exhibit regular theta oscillations during active navigation. In two pioneer studies, O'Keefe and Recce (1993) and Skaggs and collaborators (1996) observed tight correlations between theta oscillations and the sequential triggering of place cells—a phenomenon now known as "theta phase precession." Foster and Wilson (2007) then demonstrated that theta phase precessions were time-locked to the activity of corresponding place cells—a phenomenon they dubbed "theta sequences"—which essentially demonstrates that place cells not only code information about specific locations but also about the sequence in which such locations appear, given a certain learned path. Remarkably, Foster and Wilson also discovered that once a rat had learned a path, and was placed in the starting position to run the maze anew, a recap of the relevant theta sequence was recorded, albeit it occurred very quickly—less than 100 ms. In other words, the same neurons that were active during the rat's navigation were quickly reactivated, in the same order in which the relevant locations were to appear, indicating anticipatory activity of the path the rat was about to take (see Buckner, 2010, for further elaboration). Convergent evidence for anticipatory activity in the hippocampus comes from a related study by Johnson and Reddish (2007) in which rats were trained on a T-based decision maze, while cell ensembles in CA3 were recorded. As expected, they found local spatial mappings coding for specific locations in the maze. But then they used a directionally unbiased algorithm to reconstruct the neuronal activity during points at which the rat had to make a directional decision, and found, at that precise moment, a transient replay of the forward spatial mapping, indicative of a future position in the path rather than the actual position.

The number of studies reporting common engagement of neural structures underlying episodic memory and future thinking has risen steeply since 2007 (De Brigard & Gessell, 2016; Schacter, Addis, Hassabi, Martin, Spreng, & Szpunar, 2012), consolidating the view that when we remember our past and imagine what may happen in our future we deploy a common core brain



network that significantly overlaps with the so-called *default network* (Schacter, Addis, & Buckner, 2007; Spreng & Grady, 2010; Spreng, Mar, & Kim, 2009;). All such evidence supports W&D's functional thesis to the extent that the neural structures associated with episodic memory and episodic anticipation seem to overlap. But W&D's functional thesis requires more than this, as it claims not only common engagement of cognitive processes but also redeployment of the *same* information used for remembering when constructing future thoughts. Thankfully, evidence in support of this second part of the claim is also available.

Remembering past autobiographical events appears to require the deployment of a large brain network—likely the default network—of which the hippocampus and adjacent MTL areas are just one set of nodes. Of critical importance is also the reinstatement, at the time of retrieval, of the sensory-cortical regions engaged in during encoding. Initial evidence of this cortical reinstatement came from a study by Wheeler, Petersen, and Buckner (2000) in which participants were asked to study words that were paired either with pictures or with sounds while undergoing fMRI. While still in the scanner, participants received a memory test in which they were shown a word and were instructed to retrieve the associated picture or sound. Wheeler et al. discovered that pretty much the same peak cluster in the auditory cortex activated during the encoding of the word–sound association, came online during retrieval when cued with a word associated with a sound. Likewise, when the association was between a word and a picture, the same peak cluster of occipital activity registered during encoding was reactivated at retrieval. Since then, a number of neuroimaging studies have generated further support for the hypothesis that the retrieval of episodic memories requires re-activating modality-specific cortical regions activated during encoding (for a recent review, see Danker & Anderson, 2010).

A piece of neuropsychological evidence lends further support to the sensory reactivation hypothesis. Rubin and Greenberg (1998) reported the case of 11 subjects with focal lesions in the occipital cortex. Individuals with occipital cortex damage are typically described as having no memory impairment. However, because their damage prevents them from visually encoding pictorial information, assessments of their memory are usually limited to verbal tasks. But Rubin and Greenberg (1998) developed a strategy to assess the perceptual richness of their episodic memories, as compared to healthy controls, during free recall, and their studies indicated that despite being able to retrieve the gist of past experienced events, individuals with occipital damage retrieved episodic memories that were impoverished and devoid of sensory (particularly visual) details (Greenberg, Eacott, Brechin, & Rubin, 2005). Along with the neuroimaging evidence mentioned above, these results give further support to the claim that

the same sensory areas engaged during the encoding of a memory are reactivated when it is retrieved.

Still, this does not necessarily mean that the same information processed by an area at encoding is also processed at retrieval. For all we know, the same area may play different roles at each time (Bergeron, 2007). But in a recent study, Barron, Dolan, and Behrens (2013) trained participants to associate certain symbols with specific goods (“known goods,” e.g., avocado, tea) while undergoing fMRI. Next, participants were asked to imagine “novel goods,” which they had never encountered, and which consisted in the combination of two known goods (e.g., avocado tea). They hypothesized that if an imagined novel good involved the combination of two known goods, then they would see reduced activity in the voxels associated with the relevant component known goods as opposed to irrelevant known goods.<sup>2</sup> This is exactly what they found, leading Barron and colleagues, in the spirit of W&D’s thesis, to conclude that when people imagine a possible future experience they redeploy the same sensory information from encoded memories.

The evidence from the mental time travel and the sensory reactivation literatures lends credence to W&D’s functional claim, as they suggest that the neural and cognitive mechanisms associated with our capacity to remember our personal past are redeployed to use stored information to construct anticipations of possible future events. Only one piece is missing: these anticipations are supposed to be *useful*. Usefulness is a normatively laden notion, but some measures are likely proxies. Weiler, Suchan, and Daum (2010) asked participants to imagine possible events that may occur in their somewhat immediate future and to rate the *perceived plausibility* of such events while undergoing fMRI. Using ratings of perceived plausibility as parametric modulators, they discovered that hippocampal activity co-varied with perceived plausibility. In a similar vein, De Brigard, Addis, Ford, Schacter, and Giovanello (2013) used spatio-temporal partial least square analyses of fMRI data elicited during episodic recollection or episodic counterfactual thinking, and discovered that subjectively implausible counterfactual thoughts were least similar to patterns of brain activation associated with episodic recollection, while likely counterfactual scenarios recruited default network regions to a much greater extent. Previous research has also shown that during theory of mind tasks, certain hubs of the default network are

---

2. This method capitalizes on a well-known neural phenomenon known as “repetition suppression,” which refers to the reduction of neural activation for repeated stimuli relative to novel stimuli. In fMRI, the reduction of blood-oxygen-level-dependent (BOLD) signal for repeated relative to non-repeated stimuli have been widely used to study stimulus and/or feature selectivity (Grill-Spector & Malach, 2001).

modulated by considerations of *social relevance* (Krienen, Tu, & Buckner, 2010; Mitchell, Macrae, & Banaji, 2006). Using a combination of reported importance, familiarity and personal closeness, De Brigard, Spreng, Mitchell, and Schacter (2015) conducted an fMRI experiment in which participants were asked to engage in hypothetical thinking about people they perceived as unfamiliar and socially irrelevant, people they considered familiar and socially relevant, or about themselves. Default network activity was evident during simulations involving the self and socially relevant others, but not when participants imagined possible scenarios involving people they didn't know or care much about.

Assuming that perceived plausibility and social relevance are, to some extent, proxies of potential usefulness, the evidence indicates that mental simulations that have more chance of being useful tend to engage the default network to a greater extent than simulations that are perceived as less plausible or less socially relevant. In sum, the empirical evidence reviewed in this section lends strong support to W&D's functional thesis, according to which our memory system not only process information about past events but it also employs this information to construct useful anticipations of possible future events.

### 3. The Computational Thesis

According to W&D's *computational* thesis, memory must be sensitive to the statistical regularities of the information it encodes, constraining the information searched during retrieval in a probabilistic manner, in accordance with individual limitations and current goals. Thus expressed, the computational thesis shares a strong family resemblance with John Anderson's Adaptive Control of Thought, or ACT, framework (Anderson, 1990). According to ACT-R (the "R" is for "Rational"), remembering is a cognitive operation whose adaptiveness is best captured by *rational analysis*. The basic assumption is that there is always some cost associated with retrieving a memory, and that such cost is offset by the gain attained when retrieval is successful. As such, an adaptive memory system would search for a particular memory as long as the probability of recovering it, given current needs, is greater than the costs of its retrieval. The ACT-R model captures this insight in Bayesian terms: let  $H$  be the hypothesis that a particular memory ( $i$ ) is needed during a particular context, and let  $E$  be the evidence for an element of said context. Then,  $P(H_i|E) \approx P(E|H_i) P(H_i)$ , where  $P(E|H_i)$  determines the likelihood ratio that  $E$  is the case given  $H_i$  (i.e., the *contextual factor*), and  $P(H_i)$  gives the prior probability that a particular memory will be needed (i.e., the *history factor*).

Details aside (cf., Anderson & Milson, 1989; Anderson & Schooler, 2000; Schooler & Anderson, 1997), what concerns us here is that the probability of

retrieving a memory is determined by a combination of two factors: the contextual factor and the history factor. The contextual factor attempts to capture known constraints imposed by the conditions of retrieval on the probability of successfully recovering a memory. Consider the “encoding specificity principle” (Tulving & Thomson, 1973), according to which the probability of successful retrieval is increased if the information and contextual conditions present during encoding are also available during retrieval. Over the years, the encoding specificity principle has been proven in numerous proprioceptive and environmental contexts—such as studying the material underwater or while sitting in a dentist’s chair—and with all sorts of materials: pictures, words, sounds, and so on. In addition, other features of the context of retrieval may influence the probability of successfully retrieving a target memory, including attention and current interests and goals. After all, even under fully reinstated conditions of recall, if one’s mind is totally distracted, or if one’s interests or goals lie elsewhere, retrieval will be hampered. Given the multiplicity of elements present in a retrieval context, Anderson and Milson (1989) suggest that it’s better to understand the likelihood ratio as representing the context factor as the multiplicative product of all the likelihood ratios for every element of the context given  $H_i$ . As a result, certain contextual elements are better cues than others (i.e., representing a larger positive contribution to the overall product), as it is the case with elements reinstated from encoding to retrieval.

While the contextual factor is captured by the likelihood ratio that a certain  $E$  is the case given  $H_i$ ,  $P(E|H_i)$ , the history factor is captured by the prior probability that a particular memory would be needed at a time,  $P(H_i)$ . According to ACT-R, and in agreement with W&D’s computational hypothesis, this prior probability depends on the individual’s history of previous experiences. Originally, Anderson and Milson (1989, p. 705) noted that determining the history factor could be daunting, if not impossible, as one “would have to follow people about their daily lives, keeping a complete record of when they use various facts [and] such an objective study of human information is close to impossible.” As an alternative, Anderson and Schooler (1991) suggest that prior probabilities can be extracted from the statistical distribution of existent databases that would capture “coherent slices of the environment.” In an environmental database containing two years’ worth of word usage in *New York Times* headlines, they found that the odds of using a particular word in a headline was inversely correlated with its having occurred in a previous headline, with the probability decreasing the more time had passed since its last usage. Importantly, Anderson and Schooler (1991) showed that this model could capture two well-known memory retrieval effects: the recency effect (Murdock, 1962)—whereby people who study lists of items remember items presented toward the end better than those presented

in the middle—and the word-frequency effect (Gregg, 1976)—whereby high-frequency words are remembered better than low-frequency words. Taken together, the context and the history factors indicate that the probability that a certain memory will be needed in a particular context can be predicted from the probability that it has been needed in the recent past in relevantly similar contexts.<sup>3</sup>

Despite ACT-R's impressive results, it seems clear that using priors based on statistical distributions of limited “slices of the environment” does not quite capture the sense in which W&D meant that the history of *the individual* guides the probabilistic reconstruction of the retrieved memory. Indeed, their suggestion was to liken the individual's history to the previously acquired knowledge one brings to bear at the time of retrieval, which—they suggest—may be *schematic* in the sense introduced by Bartlett. Research on the effects of schemas on memory has tended to focus on two issues. The first issue concerns the fact that schematic knowledge increases recognition of schema-inconsistent information relative to schema-consistent information (Bower, Black, & Turner, 1979; Rojahn & Pettigrew, 1992). For instance, in a list of 20 words of which 18 are categorically related and 2 are complete outliers, the probability of remembering those two is higher than the probability of recalling any of the category-consistent ones—provided one controls for serial position effects, frequency, etc. The second issue concerns the effect of schema-consistent relative to schema-inconsistent information on false alarms. For example, in a study by Brewer and Treyens (1981), participants were asked to wait in an office for a few minutes while the experimenter came to fetch them. When the experimenter came back, participants were moved to a different room, where they received a surprise recognition test asking them to recall whether or not items on a list were present in the office they were just in. The list included both “old” items—that is, items that were indeed in the office—and “new” items. Critically, some of the new items—the “lures”—corresponded to objects one would typically find in an office, while the other new items were elements one would rarely find there. Brewer and Treyens (1981) found that participants were much more likely to falsely recognize office-consistent lures than office-inconsistent new items, a finding that has been replicated, in various forms, over the last few decades (Lampinen, Copeland, & Neuschatz, 2001). Many

---

3. In the past, I suggested that this kind of model dovetails with Andy Clark's (2013) hierarchical predictive approach, as the context and history factors can be combined in a hierarchical model that tries to find the most probable memory—that is, that which minimizes prediction error—for a needed memory given a certain cue. In other words, I suggested that the ACT-R-based models can be read as describing how memory retrieval attempts to minimize prediction error when finding the optimal memory given the costs of its retrieval and the organism's current needs (De Brigard, 2012). I do not pursue this line here further, but see Lin (2015).

variations on this general finding strongly suggest that schematic knowledge increases false alarms to schema-consistent relative to schema-inconsistent items.

From the Bayesian perspective, these results suggest that false alarms tend to occur for items with high prior probabilities, given a certain sample, while memory correctly rejects items with a low prior probability. What is less clear is the effect of schematic knowledge on “hits”—that is, correctly recalled items. In a beautiful experiment that speaks to this question, Huttenlocher, Hedges, and Duncan (1991) showed participants images, for just one second, of a black dot in a white circle with a little dot in the center, and asked them to reproduce the location of the black dot after only two seconds of retention interval. They discovered that accuracy in reproducing the location of the dot depended on two reference strategies that participants, unwittingly, brought to bear in the test. The first helped them code the location of the dot in relation to an imagined small circumference around it (“the fine-grain level”). The second helped them code the location of the black dot as falling within an angular sector determined by a radial-polar coordinate system—tantamount to a slice on a pizza pie (“the coarse-grain level”). Huttenlocher et al. demonstrated that black dots presented univocally within well-delimited fine- and coarse-grain locations were more accurately reproduced than black dots presented in locations that were not clearly delimited for either or both of the reference strategies (e.g., black dots presented in what would have been a boundary between imaginary pizza slices, or farther from both the center and the outer circumference, where there are no clear reference points to locate the fine-grained imaginary circle). Moreover, they found that the degree of error could be calculated with the same strategy. Huttenlocher et al.’s model suggests that people use prior topological knowledge to make predictions of where a black dot may be presented, and that black dots falling within the predicted location are better retained. Conversely, there is a monotonic decrease in accuracy linearly related to the degree to which the presented black dot deviates from the predicted location.

Inspired by this line of research, Steyvers and Hemmer (2012) conducted a study looking at the effect of prior knowledge of naturalistic scenes on hit and false alarm rates during a memory test. They employed a two-stage method: first, in a norming session, participants were asked to list the objects they would expect to see in a certain scene (e.g., an urban scene) and to list the objects they could see in a picture of that scene; second, a different group of participants were shown ten images from the set used in the norming session, one at a time, for either 2 or 10 seconds, and they were immediately asked to recall as many items as they could. Consistent with Huttenlocher et al.’s (1991) observations, Steyvers and Hemmer discovered that items that were easily named as belonging to a scene-category and more frequently found in a picture of the relevant scene were better

remembered than items that were less easily named and/or less frequently mentioned during the norming session. In fact, for the first four or five items freely recalled during testing, whether participants have seen the scene for 2 or 10 seconds did not make any difference. The combined prior probability of finding an item in scene of a certain type—tantamount to the “coarse-grain level—and of finding an item in a particular instantiation of a certain scene type—akin to the “fine-grained level”—was an equally good predictor of hit rates regardless of whether the encoding time was 2 or 10 seconds.

While Steyvers and Hemmer’s strategy (i.e., selecting priors from normed responses given by individuals in the same cohort as the experimental subjects; see Hemmer & Steyvers, 2009) represents an advance over ACT-R’s “slice of the environment” approach, it still does not capture the sense in which W&D speak of individuals’ memory being constrained by *their own* history and previous knowledge. Anderson and Milson’s (1989) concern still looms large: acquiring a schema takes time, and it is not easy to track the process of acquiring a schema to further examine it during a memory test in the controlled environment of the experimental laboratory. For that reason, most experiments aimed at directly assessing how memory performance is affected by differences in acquired schemas involve comparing within-subject performance for different pre-acquired schemas (Graesser & Nakamura, 1982; Roediger & McDermott, 1995) or between-subject performance for individuals with different schematic expertise (e.g., Arkes & Freedman, 1984; Chase & Simon, 1973; de Groot, 1966). An example of this last strategy is a study by Castel, McCabe, Roediger, and Heitman (2007) in which football experts and nonexperts were compared in a recognition test of animal names, some of which were also names of football teams. Castel and colleagues found that, if the animal names also happened to be names of football teams, football experts were better at recognizing previously seen animal names relative to non-experts, but also had more false alarms to animal name lures. Attributing these differences to the participant’s schematic knowledge of football is reasonable, yet this kind of experimental paradigm does not allow the direct manipulation of schema acquisition to study their effect on memory performance at the individual level.

Recently, however, researchers have started to explore a different strategy, based on growing evidence that schema-acquisition may be an instance of category learning (Davis, Xue, Love, Preston, & Poldrack, 2014; Love, 2013; Sakamoto, 2012; Sakamoto & Love, 2004). In one study, Palmeri and Nosofsky (1995) asked participants to learn to classify 16 geometric stimuli according to a simple rule. Although most stimuli fit the rule, the learning list included exceptions. A subsequent recognition test showed that subjects recognized category-inconsistent items at a higher rate than category-consistent ones. This recognition advantage



for rule-inconsistent relative to rule-consistent exemplars parallels the aforementioned recognition advantage for schema-inconsistent versus schema-consistent items. But the parallels do not end here. In a meta-analysis on schema-dependent recognition memory, Rojahn and Pettigrew (1992) report that the recognition advantage for schema-inconsistent information increases as the proportion of schema-inconsistent to schema-consistent items becomes smaller, a result that was later replicated by Sakamoto and Love (2004) in a category-learning task. And since category-learning can occur relatively quickly, researchers interested in the role of schematic knowledge on recognition memory are starting to employ category-learning paradigms to explore the role of schematic knowledge on recognition memory.

Using a novel category-learning paradigm, De Brigard, Brady, Ruzic, and Schacter (2017) studied the effect of acquiring new schematic knowledge on memory retrieval. To that end, they first trained participants to categorize computer-created flowers that varied among several features (e.g., color of petals, shape of center, etc.) A critical value was randomly determined as the category-inclusion criterion (e.g., red petals), and it was sampled 50% of the time during the learning period. Unbeknownst to the subject, there was another nonlearned category determined by another random feature (e.g., yellow center) that was also sampled 50% of the time. Once they learned the category, participants would study a set of flowers, and would then receive a recognition test. De Brigard et al. found that having learned a category increases both hit and false-alarm rates during recognition relative to items that did not belong to the learned category. However, items from the nonlearned category were also better remembered and elicited more false alarms than items that did not belong to either the learned or the nonlearned categories. Following category-learning paradigms, this suggests that people are more likely to correctly recall but also to false alarm to items that are more frequently encountered during learning than to those that are less frequently encountered. However, in a follow-up experiment controlling for the frequency of presentation of each feature during learning, De Brigard et al. showed that when all features are equally sampled, the effect on hit rates goes away, but the false alarm effect for items of the learned category remains, suggesting—just as Anderson and the ACT-R model predicted—that the frequency of prior encounters of a relevant item is only one factor affecting retrieval. Contextual factors, such as current goals and attentional allocation, play also a critical role.

This selective review of recent results in computational modeling and cognitive psychology suggests that the idea of treating memory retrieval as a probabilistic process resulting from both the frequency of prior experiences as well as the contextual factors of retrieval, is feasible and explanatorily powerful. Although the picture is still incomplete, I believe these lines of evidence converge



on W&D's computational thesis, according to which the process of retrieving a memory is probabilistic, and depends on priors determined by the statistical regularities of the relevant information it encodes as well as the individual's limitations and current goals.

#### 4. The Metaphysical Thesis

In section 2, I reviewed evidence coming from cognitive psychology, neuroscience, and neuropsychology to support W&D's claim that our (episodic) memory system processes information about the past and mines this information to construct useful anticipations of possible future events. In section 3, I reviewed several lines of evidence from computational psychology and cognitive science that support W&D's claim that the statistical regularities of the information we encode, which along with individual's limitations and current goals, constrain the informational space searched during memory retrieval in a probabilistic manner. In this last section, I argue that these results accord with W&D's contention that memories do not exist as subpersonal brain structures encoding particular contents but as person-level psychological phenomena describable from the intentional stance. Moreover, I argue that the functional, computational, and metaphysical theses form a coherent and strong view on the nature of episodic memory.

Recall W&D's remark that for a past event to make a difference now it must leave a long-term trace with the potential to become operational when it is needed. In the spirit of Dennettian reverse-engineering, let's ask about the nature of memory traces, given the evidence we have marshaled in support of the functional and computational theses. One possibility, rooted in contemporary analytic philosophy, is to consider memory traces as subpersonal beliefs encoded in a language of thought (Fodor, 1975). Surprisingly, despite four decades of criticism, appeals to the language of thought to explain the nature of memory traces are alive and well in philosophy. For instance, Bernecker (2008, 2010) has argued that memory traces are dispositional explicit beliefs whose intentional content refers to an experienced event. This definition needs some unpacking. First, a dispositional—as opposed to occurrent—belief is a belief that one held antecedently but that is currently unarticulated consciously (Bernecker, 2010, p. 84). Second, the content of the belief is explicit, as opposed to implicit, if the representation carrying the content of the belief is “actually present in [one's] mind in the right sort of way, for example, as an item in your ‘memory box’ or as a trace storing a sentence in the language of thought” (Bernecker, 2010, p. 29). Thus, from Bernecker's perspective, *S* remembers that *p* if and only if there is a subpersonal explicit belief *S* once held, the content of which is represented in

some stable trace in *S*'s brain, which has the disposition to become occurrent during conscious retrieval.

Unfortunately, much of what we know about the cognitive psychology and neuroscience of episodic memory speaks against this subpersonal belief view of memory traces. First, the subpersonal belief view assumes that memory beliefs perdure, unchanged, across four discrete and independent stages: belief formation, belief encoding, consolidation or "storage in some 'memory box'" (to use Bernecker's expression), and retrieval and conscious articulation. But contrary to the subpersonal belief view, these processes are neither discrete nor independent. To make this point clear, consider an ordinary experience: you are driving your car, when all of the sudden the car in front of you fails to stop at an intersection and gets hit by an incoming truck, all while you maneuver to avoid the collision. The whole event takes no more than 10 seconds. A few minutes later a policeman arrives and asks for your witness testimony, which requires you to remember the event.

As you may expect, during the experience of the event you would probably allocate attention to only a subset of elements at the scene (e.g., the wheel, your feet), and for many of them, your attention would only be transiently allocated, if at all (e.g., the grass on the curb, the traffic sign the driver missed). A wealth of psychological evidence shows that retention of episodic information depends on it having been attended during encoding, as attention is considered necessary for conscious perception (De Brigard & Prinz, 2010), and conscious perception of a stimulus is required for its successful encoding (Craik & Tulving, 1975; for a review, De Brigard, 2011). But this does not mean that attention precedes conscious perception, or that perception precedes encoding. In fact, neural evidence suggests that a number of attention-dependent neural processes occur during conscious perception but persist during encoding, even after attention has been shifted away—e.g., neuronal depolarization (Jensen & Lisman, 2005), sustained spiking (Fransen, Alonso, & Hasselmo, 2002; Hasselmo, 2007). A great deal of conscious perception and memory encoding occurs in parallel, and attempts at segmenting the processes into sequential stages are artificial, and do not reflect the psychological complexity of these processes.

Alas, this artificiality is assumed in the subpersonal view of memory representations, as it supposes that there is a particular moment in which the content of a belief is formed, and a subsequent moment in which that very content is encoded. Consider the experience of avoiding the car collision. At what point did you finalize the process of forming a belief about the experience of avoiding the car collision? Did you immediately encode it? When? When you stopped thinking about it? What if you didn't stop thinking about it until the police asked for

your testimony? Would that mean that you kept forming the belief but hadn't gotten around to encoding it? There are causal processes underlying the perception of the collision-avoidance event in virtue of which the experience is encoded, but the level of description of these casual processes does not correlate directly with the intentional level of description as is assumed by the sub-personal view of memory representations. The idea of a transparent mapping between intentional descriptions and discrete neural processes is psychologically unrealistic.

An even more pressing difficulty pertains to the moment in which encoding ends and consolidation occurs. According to the so-called standard model of consolidation, while encoding requires the interaction of the hippocampus and modality-specific areas in the neocortex, at some point (which can vary from hours to days) the hippocampus is no longer necessary for the maintenance of the memory representation, which in turn becomes stable—consolidated—as a stand-alone neural network poised to be retrieved by the interaction of a cue and the prefrontal cortex (McClelland, McNaughton, & O'Reilly, 1995; Squire, 1984). The standard model accounts for three pieces of evidence from individuals with hippocampal damage: their temporally graded retrograde amnesia (i.e., Ribot's law), their profound anterograde amnesia but preserved short-term memory, and the fact that both semantic and episodic memory are equally affected. In addition, this standard model of consolidation "at the systems level" is thought to dovetail with theories of consolidation at the synaptic level. The prevalent view on synaptic consolidation holds that experiences are encoded as changes in connectivity among the neurons originally involved in processing the perceptual information. From this perspective, learning consists in the activation and reactivation of neural networks whose co-activation strengthens their connection weights until they become highly selective for their proximal stimulus.<sup>4</sup> Details aside, the moral of the standard model is that experiencing an event, such as the avoidance of the car collision, involves the engagement of several regions of modality-specific sensory cortices: the auditory cortex would process auditory information, such as the sound of the cars crashing; the visual cortex would process visual information, such as the colors and shapes you see through the windshield; the lateral temporal cortices would probably help to categorize the perceived objects on the street, and so on (Frankland & Bontempi, 2005). During this process, the hippocampus—presumably modulated by the

---

4. First articulated by Hebb (1949), this view has found support in molecular and genetic neurobiology (e.g., Kandel, 1976; Silva, Kogan, Frankland, & Kida, 1998) as well as computational neuroscience (McClelland & Goddard, 1996). The precise mechanisms of these neural networks are complex, and involve a number of processes, such as enzymatic production (Silva, Stevens, Tonegawa, & Wang, 2002), gene regulation (Kida et al., 2002), and the formation of novel dendritic spines (Engert & Bonhoeffer, 1999).

frontoparietal attentional network—is binding together these cortical areas into a larger hippocampal-neocortical network (McClelland et al., 1995). When binding is no longer required, a memory trace is said to be “consolidated” in the neocortex, in the form of a stable neuronal assembly ready to be reactivated by the prefrontal cortex given the appropriate cue.

However, three recent lines of scientific evidence suggest that the standard model is, at best, inaccurate. First, a meta-analysis of individuals with MTL amnesia found that their retrograde amnesia for detailed autobiographical events extends for decades, sometimes even for their lifetimes, whereas the retrograde amnesia for semantic memory is less extensive, temporally graded, and differentially compromised depending on whether it involves public events, world facts or vocabulary (Nadel & Moscovitch, 1997). Indeed, the degree of retrograde amnesia is directly proportional to the amount of hippocampal damage, and different regions of the MTL differentially contribute to the formation of episodic, spatial, and semantic memories (Moscovitch et al., 2005; Nadel, Samsonovitch, Ryan, & Moscovitch, 2000). These results indicate not only that the hippocampus may actually be required during both retrieval and encoding (Eldridge, Knowlton, Furmanski, Bookheimer, & Engel, 2000; Ryan, Cox, Hayes, & Nadel, 2008), but that some memory representations may never reach a point in which they are independent of the hippocampus, and thus consolidated in the standard sense.

The second line of evidence speaks against the idea that, once consolidated, memory representations remain stable and unchanged. A number of studies in animal neurobiology have shown that amnesic interventions that are effective during the initial consolidation of an event are also effective when the supposedly consolidated memory is reactivated, suggesting that the act of retrieving a memory renders its content labile and prone to modifications (Hardt, Einarsson, & Nader, 2010; Nader & Einarsson, 2010). Indeed, the view that memories can be updated and modified during reactivation, as opposed to being stable and unchanging, is becoming the received account of why people are likely to misremember a plausible event as having occurred if they previously imagined it (imagination inflation; Garry, Manning, & Loftus, 1996), to wrongly recognized as experienced information that was misleadingly introduced at the time of retrieval (post-event misinformation; Loftus & Hoffman, 1989), and to false alarm to lures that are conceptually or semantically related to studied items (Roediger & McDermott, 1995). Many interpret these results as suggesting that retrieval renders memory traces liable to distortion, as information processed online while one is remembering can infiltrate and/or modify the re-encoded memorial content.

Finally, and as mentioned in section 2, extant neuroscientific evidence suggests that there isn’t a dedicated storage unit for episodic memories. At best, “storage” is a metaphoric label for the following dynamic process: first, connections

in the cortico-hippocampal neural network that is engaged during the initial experience of the event are strengthened; second, something like an “index” is formed—likely in the hippocampus—indicating which cells of the distributed neuronal assembly engaged during the encoding of the experience need to be reactivated when remembering takes place (Moscovitch et al., 2005); finally, retrieval consists in the reinstatement of the pattern of activation the brain was in during encoding—a reinstatement that, because of the dynamic, labile and changing process of “storage,” is never identical to but is close enough to its original form. Contrary to the subpersonal belief view of memory representations, the evidence speaks against there being a trace in the brain that, upon encoding, lies dormant, patiently carrying the intentional content of the encoded experience until it is retrieved. Indeed, the very same regions that were engaged during encoding and are recruited later on during retrieval are constantly redeployed in the interim to serve all sorts of roles instead of simply being dedicated to encoding one particular experience.

How can we reconcile all this evidence against the subpersonal belief view of memory with W&D’s own remark that, for a past event to make a difference now, it needs to leave a long-term trace with the capacity to become useful when needed? The answer, I surmise, hinges on the term “capacity,” for what needs to be stored is not the intentional content of a memory per se but, rather, the disposition to entertain such intentional content when it is needed. W&D are clear about this point: being able to retain the “ability to encode useful information and to decode it in precisely those circumstances where it can be useful” (Westbury & Dennett, 2000, p. 24) is, indeed, a process dependent on subpersonal mechanisms. But this sense of storing information in the brain should not be conflated with the ordinary, personal-level concept of memory (Westbury & Dennett, 2000). To remember that *p*, then, is not to possess a subpersonal memory belief carrying the relevant intentional content from encoding to retrieval, but to exhibit the kind of behavior that is optimally described and predicted—even for the individual herself—by ascribing the memory that *p* from the intentional stance.

And here is the great payoff from this austere metaphysical view of memory traces: it dovetails nicely with the evidence marshaled in favor of the functional and computational theses. Instead of evolving to store artificially divided temporal slices of experienced events, our predictive brains developed the capacity to instantiate the dispositional property of reinstating the state they were in during encoding at the time of retrieval. The computational savings of this maneuver are enormous: imagine the amount of storage—the size of the “memory box”—needed for your average-size long-term memory. But our brains are also dynamic machines, engaged in all sorts of exchanges with the external and internal

environment that lead to constant changes in its neuronal connections. Given these changes, reinstating the precise state the brain was in during encoding at the time of retrieval may actually be impossible. The next best solution is to maximize the chances of reinstating this initial state by constraining the reactivation probabilistically, in the sense discussed here. Think, for instance, how easily this view can fit the phenomenon of encoding specificity (Craig & Tulving, 1975). If you help the sensory cortices reinstate the situation they were in, during encoding, at the time of retrieval, by stimulating them with the same stimuli and/or in the same proprioceptive and environmental context in which encoding occurred, this immediately increases the probability of getting the rest of the relevant neural network reactivated. And, finally, I suggest that these very same probabilistic constraints help to explain why the brain would mine the past to construct mental simulations of possible future events. If remembering is the process by means of which we reconstruct the most likely past given the evidence available at the time of retrieval, then why not employ the exact same machinery to generate anticipations of what may come? It is very unlikely that the future would be exactly as the past was, but it is definitively more likely that it would be as the past could have been. The flexibility afforded by the probabilistic reconstruction constrained by prior experiences is not only an optimal strategy to reconstruct the past but also to predict the future.

## 5. A (Personal) Conclusion

There is a sense in which this chapter could be seen as an updated reading of perhaps the only paper Dan Dennett has written on episodic memory as its main topic. After all, I have presented my argument as an attempt to demonstrate how W&D's functional, computational, and metaphysical theses, not only find strong support from recent scientific developments but also form a coherent and promising way of understanding memory from a philosophical point of view. But there is another sense in which the current paper is not so much an investigation into how Dennett reverse-engineers memory as much as it is an exercise in how I reverse-engineer my *own views* on memory. It is evident there was much more wishful reading than there was exegesis in this exercise. And yet, even if mine, I think the view offered here is strictly Dennettian. You see, Dan Dennett shares with Gilbert Ryle, his adviser, an enviable and profoundly generous mentoring trait: a way of influencing one's views without imposing, a way of lovingly letting you discover the advantages of seeing things from his perspective without ever belittling your own. Recently, Dennett articulated this experience in an anecdote, included in the foreword (Dennett, 2015) to a recent volume on his first book, *Content and Consciousness* (Muñoz-Suárez & De Brigard, 2015). In

talking about the disorienting experience of working on his dissertation—which would later become that first book—under the direction of Ryle, he mentions feeling that Ryle never “fought back” but instead tended to agree with his good points, pressing him on mere adjustments here and there. Dennett even confesses to thinking that he hadn’t learned much philosophy from Ryle. But, then, he recalls:

I finished a presentable draft of my dissertation in the minimum time (six terms or 2 years) and submitted it, with scant expectation that it would be accepted on first go. On the eve of submitting it, I came across an early draft of it and compared the final product with its ancestor. To my astonishment, I could see Ryle’s influence on every page. How had he done it? Osmosis? Hypnotism? This gave me an early appreciation of the power of indirect methods in philosophy. You seldom talk anybody out of a position by arguing directly with their premises and inferences. Sometimes it is more effective to nudge them sideways with images, examples, and helpful formulations that stick to their habits of thought (p. vii).

Looking back at my own experience with my doctoral dissertation, I can’t help but think that I, too, had been “Ryled” by Dennett. Despite his being an external member of my dissertation committee, I had the great fortune of being able to share my thoughts with Dan often as I was writing. At no point did he mention the paper he wrote with Westbury. I learned about this paper only after I had sent him the last draft of my dissertation—he mentioned the W&D paper in passing, more interested, I thought, in telling me how the paper came to be than about its contents. And so, I put it aside, and only came to read it much later on, even after two of the chapters of my dissertation were already on their way to being published (e.g., De Brigard, 2014a, 2014b). In a way, I felt terrible for not having read it before, as I should have, for after I did, I, too, came to see Dennett’s influence on every page I wrote. But perhaps that was exactly Dan’s intention. Perhaps that was his indirect way of showing me the advantages of adopting a Dennettian view on memory. I hope this chapter helps to begin an articulation of such a view, and I also hope it allows me to gratefully express how much I’ve learned from Dan.

## Acknowledgments

Thanks to Kourken Michaelian, Bryce Huebner, and Gregory Stewart for useful comments on an earlier draft.



## Works Cited

- Addis, D. R., Wong, A. T., & Schacter, D. L. (2007). Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration. *Neuropsychologia*, 45, 1363–1377.
- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.
- Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, 96, 703–719.
- Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, 2, 396–408.
- Anderson, J. R., & Schooler, L. J. (2000). The adaptive nature of memory. In E. Tulving & F. Craik (Eds.), *The Oxford handbook of memory* (pp. 557–570). Oxford, UK: Oxford University Press.
- Arkes, H. R., & Freedman, M. R. (1984). A demonstration of the costs and benefits of expertise in recognition memory. *Memory and Cognition*, 12, 84–89.
- Atance, C. M., & O'Neill, D. K. (2001). Episodic future thinking. *Trends in Cognitive Sciences*, 5, 533–539.
- Barron, H. C., Dolan, R. J., & Behrens, T. E. (2013). Online evaluation of novel choices by simultaneous representation of multiple memories. *Nature Neuroscience*, 16, 1492–1498.
- Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, England: Cambridge University Press.
- Bergeron, V. (2007). Anatomical and functional modularity in cognitive science: Shifting the Focus. *Philosophical Psychology*, 20, 175–195.
- Bernecker, S. (2008). *The metaphysics of memory*. Dordrecht, The Netherlands: Springer.
- Bernecker, S. (2010). *Memory*. Oxford, UK: Oxford University Press.
- Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology*, 11, 177–220.
- Brewer, W. F., & Treyens, J. C. (1981). Role of schemata in memory for places. *Cognitive Psychology*, 13, 207–230.
- Buckner, R. L. (2010). The role of the hippocampus in prediction and imagination. *Annual Reviews Psychology*, 61, 27–48.
- Cabeza, R., & St. Jacques, P. (2007). Functional neuroimaging of autobiographical memory. *Trends in Cognitive Sciences*, 11, 219–227.
- Castel, A. D., McCabe, D. P., Roediger, H. L., III, & Heitman, J. L. (2007). The dark side of expertise: Domain specific memory errors. *Psychological Science*, 18, 3–5.
- Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed.), *Visual information processing* (pp. 215–281). New York: Academic Press.
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Science*, 36, 181–204.
- Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford: Oxford University Press.



- Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, 104, 268–294.
- Danker, J. F., & Anderson, J. R. (2010). The ghosts of brain states past: Remembering reactivates the brain regions engaged during encoding. *Psychological Bulletin*, 136(1), 87–102.
- Davis, T., Xue, G., Love, B. C., Preston, A. R., & Poldrack, R. A. (2014). Global neural pattern similarity as a common basis for categorization and recognition memory. *Journal of Neuroscience*, 34, 7472–7484.
- D'Argembeau, A., & Van der Linden, M. (2004). Phenomenal characteristics associated with projecting oneself back into the past and forward into the future: Influence of valence and temporal distance. *Consciousness and Cognition*, 13, 844–858.
- D'Argembeau, A., & Van der Linden, M. (2006). Individual differences in the phenomenology of mental time travel: The effect of vivid visual imagery and emotion regulation strategies. *Consciousness and Cognition*, 15, 342–350.
- De Brigard, F. (2011). The role of attention in conscious recollection. *Frontiers in Psychology*, 3, 29.
- De Brigard, F. (2012). Predictive memory and the surprising gap: Commentary on Andy Clark's "Whatever Next? Predictive Brains, Situated Agents and the Future of Cognitive Science." *Frontiers in Psychology*, 3, 420.
- De Brigard, F. (2014a). Is memory for remembering? Recollection as a form episodic hypothetical thinking. *Synthese*, 191, 155–185.
- De Brigard, F. (2014b). The nature of memory traces. *Philosophy Compass*, 9, 402–414.
- De Brigard, F. (2015). What was I thinking? Dennett's *Content and Consciousness* and the reality of propositional attitudes. In C. M. Muñoz-Suárez & F. De Brigard (Eds.), *Content and Consciousness Revisited* (pp. 49–71). New York, NY: Springer.
- De Brigard, F., Addis, D., Ford, J. H., Schacter, D. L., & Giovanello, K. S. (2013). Remembering what could have happened: Neural correlates of episodic counterfactual thinking. *Neuropsychologia*, 51, 2401–2414.
- De Brigard, F., & Prinz, J. (2010). Attention and consciousness. *WIREs Interdisciplinary Reviews: Cognitive Science*, 1(1), 51–59.
- De Brigard, F., Spreng, R. N., Mitchell, J. P., & Schacter, D. L. (2015). Neural activity associated with self, other, and object-based counterfactual thinking. *NeuroImage*, 109, 12–26.
- De Brigard, F., & Gessell, B. S. (2016). Time is not of the essence: Understanding the neural correlates of mental time travel. In S. B. Klein, K. Michaelian, & K. K. Szpunar (Eds.), *Seeing the future: Theoretical perspectives on future-oriented mental time travel*. New York, NY: Oxford University Press, 153–180.
- De Brigard, F., Brady, T. F., Ruzic, L., & Schacter, D. L. (2017). Tracking the emergence of memories: A category-learning paradigm to explore schema-driven recognition. *Memory and Cognition*, 45(1), 105–120.

- de Groot, A. D. (1966). Perception and memory versus thought: Some old ideas and recent findings. In B. Kleinmuntz (Ed.), *Problem solving* (pp. 19–50). New York, NY: Wiley.
- Dennett, D. C. (1969). *Content and Consciousness*. London: Routledge and Kegan Paul.
- Dennett, D. C. (1976). Are dreams experiences? *Philosophical Review*, 85, 151–171.
- Dennett, D. C. (1978). *Brainstorms*. Cambridge, MA: MIT Press.
- Dennett, D. C. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.
- Dennett, D. C. (1991). *Consciousness Explained*. New York, NY: Little, Brown.
- Dennett, D. C. (1994). Cognitive science as reverse engineering: Several meanings of ‘top-down’ and ‘bottom-up’. In D. Prawitz, B. Skyrms, & D. Westerståhl (Eds.), *Logic, methodology and philosophy of science IX* (pp. 679–689). Amsterdam, North-Holland: Elsevier.
- Dennett, D. C. (2015). Foreword to C. Muñoz-Suárez, & F. De Brigard, *Content and Consciousness Revisited* (pp. v–x). New York, NY: Springer.
- Eldridge, L. L., Knowlton, B. J., Furmanski, C. S., Bookheimer, S.Y., & Engel, S.A. (2000). Remembering episodes: A selective role for the hippocampus during retrieval. *Nature Neuroscience*, 3, 1149–1152.
- Engert F., & Bonhoeffer, T. (1999). Dendritic spine changes associated with hippocampal long-term synaptic plasticity. *Nature*, 399, 66–70.
- Fodor, J. 1975. *The language of thought*. Cambridge, MA: Harvard University Press.
- Foster, D. J., & Wilson, M. A. (2007). Hippocampal theta sequences. *Hippocampus*, 17, 1093–1099.
- Frankland, P. W., & Bontempi, B. (2005). The organization of recent and remote memories. *Nature Reviews Neuroscience*, 6, 119–130.
- Fransen, E., Alonso, A. A., & Hasselmo, M. E. (2002). Simulations of the role of the muscarinic-activated calcium-sensitive non-specific cation current I(NCM) in entorhinal neuronal activity during delayed matching tasks. *Journal of Neuroscience*, 22, 1081–1097.
- Garry, M., Manning, C. G., & Loftus, E. (1996). Imagination inflation: Imagining a childhood event inflates confidence that it occurred. *Psychonomic Bulletin and Review*, 3, 208–214.
- Graesser, A. C., & Nakamura, G. V. (1982). The impact of a schema on comprehension and memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 16, pp. 59–109). New York, NY: Academic Press.
- Greenberg, D. L., Eacott, M. J., Brechin, D., & Rubin, D. C. (2005). Visual memory loss and autobiographical amnesia: A case study. *Neuropsychologia*, 43, 1493–1502.
- Gregg, V. (1976). Word frequency, recognition, and recall. In J. Brown (Ed.), *Recall and recognition* (pp. 183–216). New York, NY: Wiley.
- Grill-Spector, K., & Malach, R. (2001). fMR-adaptation: A tool for studying the functional properties of human cortical neurons. *Acta Psychologica*, 107, 293–321.

- Hardt, O., Einarsson, E. Ö., & Nader, K. (2010). A bridge over troubled water: Reconsolidation as a link between cognitive and neuroscientific memory research traditions. *Annual Review of Psychology*, 61, 141–167.
- Hassabis, D., & Maguire, E. A. (2007). Deconstructing episodic memory with construction. *Trends in Cognitive Sciences*, 11, 299–306.
- Hasselmo, M. E. (2007). Encoding: Models linking neural mechanisms to behavior. In R. Roediger, Y. Dudai, & S. Fitzpatrick (Eds.), *Science of memory: Concepts*. New York, NY: Oxford University Press, 124–143.
- Hebb, D. O. (1949). *The organization of behavior*. New York, NY: Wiley.
- Helmholtz, H. von (1866). *Concerning the perceptions in general*. In *Treatise on physiological optics*, vol. III, 3rd edn (translated by J. P. C. Southall 1925 *Optical Society of America* Section 26, reprinted, New York: Dover, 1962).
- Hemmer, P., & Steyvers, M. (2009). A Bayesian account of reconstructive memory. *Topics in Cognitive Science*, 1, 189–202.
- Hohwy, J. (2013). *The predictive mind*. Oxford, UK: Oxford University Press.
- Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review*, 98, 352–376.
- Ingvar, D. H. (1985). “Memory of the future”: An essay on the temporal organization of conscious awareness. *Human Neurobiology*, 4(3), 127–136.
- Jensen, O., & Lisman, J. E. (2005). Hippocampal sequence-encoding driven by cortical multi-item working memory buffer. *Trends in Neurosciences*, 28(2), 67–72.
- Johnson, A., & Redish, D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *Journal of Neuroscience*, 27, 12176–12189.
- Kandel, E. R. (1976). *Cellular basis of behavior: An introduction to behavioral neurobiology*. San Francisco, CA: W. H. Freeman.
- Kant, I. (1781/1999). *Critique of pure reason*. In Guyer, P., and Wood, A. W. (Eds.), *The Cambridge edition of the works of Immanuel Kant*. Cambridge, UK: Cambridge University Press.
- Kida S., Josselyn, S. A., de Ortiz, S. P., Kogan, J. H., Chevere, I., Masushige, S., & Silva, A. J. (2002). CREB required for the stability of new and reactivated fear memories. *Nature Neuroscience*, 5, 348–355.
- Klein, S. B., Loftus, J., & Kihlstrom, J. F. (2002). Memory and temporal experience: The effects of episodic memory loss on an amnesic patient’s ability to remember the past and imagine the future. *Social Cognition*, 20, 353–379.
- Korsakoff, S. S. (1889/1996). Medico-psychological study of a memory disorder. *Consciousness and Cognition*, 5, 2–21.
- Krienen, F. M., Tu, P. C., & Buckner, R. L. (2010). Clan mentality: Evidence that medial prefrontal cortex responds to close others. *Journal of Neuroscience*, 30, 13906–13915.
- Lampinen, J. M., Copeland, S. M., & Neuschatz, J. S. (2001). Recollections of things schematic: Room schemas revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 1211–1222.

- Lin, Y. T. (2015). Memory for prediction error minimization: From depersonalization to the delusion of non-existence; a commentary on Philip Gerrans. In T. Metzinger & J. M. Windt (Eds.), *Open Mind*: 15(C). Frankfurt am Main, Germany: Mind Group. Retrieved from doi: 10.15502/9783958570719
- Llinás, R. R. (2001). *I of the vortex*. Cambridge, MA: MIT Press.
- Loftus, E. F., & Hoffman, H. G. (1989). Misinformation and memory: The creation of new memories. *Journal of Experimental Psychology: General*, 118, 100–104.
- Love, B. (2013). Categorization. In K. N. Ochsner & S. M. Kosslyn (Eds.), *Oxford handbook of cognitive neuroscience* (pp. 342–358). Oxford, UK: Oxford University Press.
- Maturana, H. R. (1970). *Biology of cognition* (BCL Report 9.0). Biological Computer Laboratory. Department of Electrical Engineering. Urbana, IL: University of Illinois.
- McClelland, J. L., & Goddard, N. (1996). Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus*, 6, 654–665.
- McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102, 419–457.
- Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*, 50, 655–663.
- Moscovitch, M., Rosenbaum, R. S., Gilboa, A., Addis, D. R., Westmacott, R., Grady, C., . . . Nadel, L. (2005). Functional neuroanatomy of remote episodic, semantic and spatial memory: A unified account based on multiple trace theory. *Journal of Anatomy*, 207, 35–66.
- Muñoz-Suárez, C. M., & De Brigard, F. (Eds.). (2015). *Content and Consciousness Revisited*. New York, NY: Springer.
- Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, 64, 482–488.
- Nadel, L., & Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinion in Neurobiology*, 7, 217–227.
- Nadel, L., Samsonovitch, A., Ryan, L., & Moscovitch, M. (2000). Multiple trace theory of human memory: Computational, neuroimaging and neuropsychological results. *Hippocampus*, 10, 352–368.
- Nader, K., & Einarsson, E. Ö. (2010). Memory reconsolidation: An update. *Annals of the New York Academy of Sciences*, 1191, 27–41.
- O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford, UK: Oxford University Press.
- O'Keefe, J., & Recce, M. L. (1993). Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus*, 3, 317–330.
- Okuda, J., Fujii, T., Ohtake, H., Tsukiura, T., Tanji, K., Suzuki, K. (2003). Thinking of the future and past: The roles of the frontal pole and the medial temporal lobes. *NeuroImage* 19, 1369–1380.

- Palmeri, T. J., & Nosofsky, R. M. (1995). Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 548–568.
- Race, E., Keane, M. M., & Verfaellie, M. (2011). Medial temporal lobe damage causes deficits in episodic memory and episodic future thinking not attributable to deficits in narrative construction. *Journal of Neuroscience*, 31, 10262–10269.
- Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 803–814.
- Rojahn, K., & Pettigrew, T. F. (1992). Memory for schema-relevant information: A meta-analytic resolution. *British Journal of Social Psychology*, 31, 81–109.
- Rubin, D. C., & Greenberg, D. L. (1998). Visual memory-deficit amnesia: A distinct amnesic presentation and etiology. *Proceedings of the National Academy of Sciences*, 95, 5413–5416.
- Ryan, L., Cox, C., Hayes, S., & Nadel, L. (2008). Hippocampal activation during episodic and semantic memory retrieval: Category production and category cued recall. *Neuropsychologia*, 46, 2109–2121.
- Sabra, A. I. (1989). *The Optics of Ibn Al-Haytham: On Direct Vision Books 1-3*. London: Warburg Institute.
- Sakamoto, Y., & Love, B. C. (2004). Schematic influences on category learning and recognition memory. *Journal of Experimental Psychology: General*, 133, 534–553.
- Sakamoto, Y. (2012). Schematic influences on category learning and recognition memory. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning*. New York, NY: Springer.
- Schacter, D. L., & Scarry, E. (Eds.). (2000). *Memory, brain, and belief*. Cambridge, MA: Harvard University Press.
- Schacter, D. L., & Addis, D. R., & Buckner, R. L. (2007). Remembering the past to imagine the future: The prospective brain. *Nature Reviews Neuroscience*, 8, 657–661.
- Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012). The future of memory: Remembering, imagining, and the brain. *Neuron*, 76, 677–694.
- Schooler, L. J., & Anderson, J. R. (1997). The role of process in the rational analysis of memory. *Cognitive Psychology*, 32, 219–250.
- Silva, A. J., Kogan, J. H., Frankland, P. W., & Kida, S. (1998). CREB and memory. *Annual Review of Neuroscience*, 21, 127–148.
- Silva, A. J., Stevens, C. F., Tonegawa, S., & Wang, Y. (2002). Deficient hippocampal long-term potentiation in alpha-calcium-calmodulin kinase II mutant mice. *Science*, 257, 201–206.
- Skaggs, W. W., McNaughton, B. L., Wilson, M. A., & Barnes, C. A. (1996). Theta Phase precession in hippocampal neuronal populations and the compression of temporal sequences. *Hippocampus*, 6, 149–172.

- Spreng, N. R., & Levine, B. (2006). The temporal distribution of past and future autobiographical events across the lifespan. *Memory and Cognition*, 34, 1644–1651.
- Spreng, R. N., Mar, R. A., & Kim, A. S. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience*, 21, 489–510.
- Spreng, R. N., & Grady, C. L. (2010). Patterns of brain activity supporting autobiographical memory, prospection, and theory of mind, and their relationship to the default mode network. *Journal of Cognitive Neuroscience*, 22, 1112–1123.
- Squire, L. R. (1984). Neuropsychology of memory. In P. Marler & H. Terrace (Eds.), *The biology of learning* (pp. 667–685). Berlin, Germany: Springer-Verlag.
- Steyvers, M., & Hemmer, P. (2012). Reconstruction from memory in naturalistic environments. In Brian H. Ross (Ed), *The psychology of learning and motivation* (pp. 126–144). New York, NY: Elsevier.
- St. Jacques, P., & De Brigard, F. (2015). In D. R. Addis, M. Barense, & A. Duarte, (Eds.), *The Wiley handbook on the cognitive neuroscience of memory*. New York, NY: Wiley.
- Szpunar, K., & Watson, J. M., & McDermott, K. B. (2007). Neural substrates of envisioning the future. *Proceedings of the National Academy of Sciences of the United States of America*, 104, 642–647.
- Talland, G.A. (1965). *Deranged memory: A psychonomic study of the amnesic syndrome*. New York, NY: Academic Press.
- Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80, 352–373.
- Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, 26(1), 1–12.
- Weiler, J., A., Suchan, B., & Daum, I. (2010). Foreseeing the future: Occurrence probability of imagined future events modulates hippocampal activation. *Hippocampus*, 20, 685–690.
- Westbury, C., & Dennett, D. C. (2000). Mining the past to construct the future: Memory and belief as forms of knowledge. In D. L. Schacter & E. Scarry (Eds.), *Memory, brain, and belief* (pp. 11–32). Cambridge, MA: Harvard University Press.
- Wheeler, M. E., Petersen, S. E., & Buckner, R. L. (2000). Memory's echo: Vivid remembering reactivates sensory-specific cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 97, 11125–11129.

## 3.2 REFLECTIONS ON FELIPE DE BRIGARD

Daniel C. Dennett

Felipe de Brigard disinters a lost paper of mine with Chris Westbury (2000) and shows, generously, how well it presages more recent developments. That is mainly due, I am sure, to Westbury's expert knowledge of the state of the field at the millennium—something I was happy to get his guidance on. I think Felipe is right that the view we expressed then “has received substantial support during the last decade and a half of scientific research.” Added to Westbury's knowledge and acumen was my perennial tactic of not going out on limbs I don't have to go out on to make my conceptual or theoretical points. The Bayes Bandwagon was already starting to rumble back then, but I, for one, wasn't ready to hop on with a wholehearted commitment. (I'm getting closer. It does seem, as Andy Clark, Jakob Hohwy [2013], and many others are declaring, with impressive support, that this is a breakthrough we've been waiting for.)

But I'm not swept off my feet just yet. There are still some issues regarding—sigh—*representation*, that haven't been clarified or solved by Bayesian approaches, so far as I can see. I try to shed a little light on that in *BBB*, but I'm well aware that there is still way too much hand-waving about the actual vehicles of information in the nervous system. I can use one of Felipe's closing points to expose part of it:

Contrary to the subpersonal belief view of memory representation, the evidence speaks against there being a trace in the brain that upon encoding, lies dormant, patiently carrying the intentional content of the encoded experience until it is retrieved. Indeed, the very same regions that were engaged during encoding and are later on recruited during retrieval are constantly redeployed in the interim to serve all sorts of roles, rather than just being dedicated to encode one particular experience.



How about this? If there *were* a language of thought, it would be constantly changing, subtly, as Bayesian improvements tuned the boundaries, sharpened the distinctions, rooted out the overgeneralizations, etc., much the way English changes, with new meanings, new words, and some words moving through obsolescence to extinction. Then if a relatively inert memory was stored away written in LOT1, when it is retrieved many years later, it will be distorted by the change in the meanings of the terms in the now current LOT2. And we seem to have a well-known example of just such phenomena: you return to childhood haunts and find everything much smaller than you remember it being, such as the furniture in your kindergarten classroom. It seems that your LOT terms for *small*, *medium*, and *large* were anchored to your body (not surprisingly) and they have enlarged their meaning as your body did. I am sure such cases can be explained in terms that are not quite so slavishly LOT-ish, but then, how language-like must a LOT be to count as a LOT? I don't think De Brigard's case against "dedicated" subpersonal mechanisms to encode one particular experience is secure yet, because of the plethora of in-between possibilities not yet canvassed and eliminated. Quine, in "Universal Library," (1989) showed how Borges's "Library of Babel," with its Vast but finite array of hexagonal air shafts lined with bookcases filled with books—all the possible books—could be replaced by a much smaller library, using just two symbols, 0 and 1, and then,

the ultimate economy: a cutback in the size of the volumes. Instead of admitting 500,000 occurrences of characters to each volume, we might settle for say seventeen. We have no longer to do with volumes, but with two-inch strips of text, and no call for half-calf bindings. In our two-character code the number of strips is  $2^{17}$ , or 131,072. The totality of truth is now reduced to a manageable compass. Getting a substantial account of anything will require extensive concatenation of our two-inch strips, and re-use of strips here and there. But we have everything to work with.

The problem, which Quine fully recognizes, is that now the library's *index* is doing almost all the work, since every good book—say, *Moby Dick*—now consists of a multi-"volumed" set of those two-inch strips, "extensively concatenated," many of them used thousands of times! Where is the book stored? In the little strips or in the combinatorially exploding "index" entry for the book? At some point an index is no longer a pointer or small set of pointers but rather a recipe for building a copy. Such recipes are very useful and very common these days. A PDF file is just such a recipe.

Back to the hippocampus: Does it "just" index memories or store recipes for reconstructing memories? Or something in-between, as yet unimagined? Imagine



a community that assigns responsibility for the “minutes” of every day to a trio of individuals. At sundown, the day’s trio meets briefly to go over their recollections of the day’s events, ironing out inconsistencies as best they can, and then disperse. Periodically, they individually rehearse their memories of the day. If, on some later date, a question arises about what transpired on July 8, all that has to be stored is the names of the day’s trio. As long as one of them is still alive, you get the account. The problem with this analogy, some may think, is that individual people have agendas, which differ from others’ agendas and change over time, so our recollection trios would be a problematic low-fidelity store of information about the past. Indeed, it would be, but so his human memory, which must rely on resources not ideally suited for hi-fi storage.

Getting Vast representational power out of not just finite but portable materials is still an unsolved problem.

## Works Cited

- Hohwy, J. (2013). *The predictive mind*. Oxford, UK: Oxford University Press.
- Quine, W. V. O. (1989). *Quiddities: An intermittently philosophical dictionary*. Universal Library. Cambridge, MA: Harvard University Press, 223–224.
- Westbury, C., & Dennett, D. C. (2000). Mining the past to construct the future: Memory and belief as forms of knowledge. In D. L. Schacter & E. Scarry (Eds.), *Memory, brain, and belief* (pp. 11–32). Cambridge, MA: Harvard University Press.

# 4.1

## REPRESENTATIONS AND RULES IN LANGUAGE

Ray Jackendoff

One thing that has always distinguished Dan Dennett's philosophy is his eagerness to get his hands dirty with the phenomena, to take things apart and see what's inside, to "turn the knobs" on the simplistic scenarios proffered by intuition pumps in the lore (Dennett, 2013). My goal here is to practice what he preaches.

The scenario whose knobs I'll be turning here is what might be thought of as the "lay cognitive scientist's" (and lay philosopher's) view of language: that sentences (one kind of mental representation) are constructed by combining words (another kind of mental representation) by means of rules. This scenario is of course an instance of the central stance of cognitive science: that mental activity can be described as computational operations over mental representations. But all too often this stance is given little more than lip service, and not much thought is devoted to the question of precisely what the mental representations and the operations on them *are*, beyond the demands of the experiment or computational model at hand.

In this chapter I want to wrestle with the nature of representations and rules in language, comparing a number of proposals in the literature to my own (Culicover & Jackendoff, 2005; Jackendoff, 2002). The outcome is a reconceptualization of how knowledge of language is encoded by the brain, with implications for cognitive science at large.

A road map: First, after a prelude about what a "mental representation" could be, we look at how words are encoded in the lexicon (or dictionary). We then examine three different ways to think about rules: the Procedural Strategy, the Declarative Strategy, and the No Rules Strategy. Turning the knobs, we explore a range of phenomena that lie between classical words and classical rules, concluding that the Declarative Strategy has a decided advantage in accounting for the

existence of such items. We then turn some different knobs, comparing fully productive rules—the classical rules, which work all the time—with partially productive rules, which work only *some* of the time. Again, the Declarative Strategy proves superior. Finally, we think briefly about how rules could be acquired, and show how the Declarative Strategy offers a more satisfying account (though still incomplete in many respects).

I recognize that my story here is somewhat technical for non-linguist audiences. Nevertheless, I encourage readers to bear with me. The details *do* make a difference.

## 1. What Is a Mental Representation?

I want to think of mental representations not as characterizations of whole-brain states (the way philosophers sometimes talk about knowing *P* as being in some particular brain state), but rather as data structures that pertain to a particular domain of brain activity. A model case is phonological structure in language. The sound structure of a word is encoded as a sequence of phonological segments, each of which can be characterized in terms of the values of a set of phonological distinctive features. For instance, the word *cat* is represented phonologically as the sequence of phones /k/, /æ/, and /t/, where /k/ in turn consists of the features [+consonant, -continuant, -voiced, +velar closure], and similarly for the other segments.

Phonological structure is a level of abstraction beyond the phonetic representation. In the context of beginning a stressed syllable, phonological /k/ is aspirated as phonetic [kʰ]. And the exact phonetic value of phonological /æ/ depends on the consonant following within the same syllable: longer (and in some dialects, a bit higher) if followed by a voiced consonant (as in *pad*), shorter (and lower) if followed by an unvoiced consonant (as in *pat*). And in turn phonetic representation itself is a sequence of discrete sounds, and is therefore is a major abstraction from the continuously varying acoustic signal.

When one hears the word *cat*, the distinctive features are not present in the acoustic signal; they must be computed on the basis of the acoustic signal. (This is one important reason why automatic speech recognition has proven so difficult.) Yet linguists consider these features “psychologically real” because of the role they play in determining the patterns of human languages. A feature doesn’t “represent” anything in the physical world, it is not a “symbol” for anything; rather, it imposes an equivalence class on heard and produced sounds that classifies them for linguistic purposes, that is, for human interests. The same goes in the syntactic domain, if not more so, for features such as Verb, Noun Phrase, and Suffix.

Just to be clear: the letter “k” *does* represent something. It *is* a symbol for something, namely the phonological segment /k/. Similarly, the *notation* [-voiced] represents something, namely the value of a feature in a hypothesized mental representation. But the phonological segment and the feature themselves do *not* “represent”—they are not symbols for anything.<sup>1</sup> They get their values only by virtue of their roles in the system of language. One might think of the phonological structure of a word as a sort of virtual object, like a melody or a baseball score or a turn (as in taking turns), or a dollar (as in Dennett’s, 2013, “How Much Is That in Real Money?”). These are all “cognitively real,” that is, they play a role in human construal of (or understanding of) the physical and social world—“institutional facts” in Searle’s (1995) terms. But they are not physically real in any useful sense. Again, the same arguments pertain to syntactic categories and syntactic features.

In addition to the varieties of data structures, an essential component of mental representation is the links *among* different spaces of data structures. The word *cat* and the non-word *smat* both have phonological structures. But *cat* is more than that: its phonological structure /kæt/ is linked in long-term memory with a syntactic category and a meaning. In the domain of meaning, the relevant mental representations differentiate concepts into objects, events, properties, places, and so on; the objects are differentiated into animate and inanimate and characterized in terms of shape and function; and so on down to the concept of a cat somewhere in the taxonomy, itself further differentiated into kinds of cats and cats one has known. The concept of a cat (or some version of it) could exist in the mind of a nonlinguistic organism such as an ape or a baby, without being linked to a phonological structure, that is, without being part of a word. Alternatively, in the mind of a speaker of another language, this concept could be linked to a different phonological structure (or, in a sign language speaker, to a gestural structure).

In other words, phonological structures and meanings are independent domains of mental representations, and the links *among* mental domains are as vitally important as the structures themselves. In particular, the space of phonological structures is useful precisely because it offers a rich domain that can be linked to concepts and thereby render them expressible. More generally, without such linkages or “interfaces,” a domain of mental representation would be thoroughly encapsulated, and hence useless to the organism in regimenting perception and determining behavior.

---

1. But they are discrete. The term *symbolic* in cognitive science has frequently conflated being discrete with being a symbol for something.

It is a habit in much of cognitive science to depict an interface as an arrow connecting two boxes. What this obscures is that this process is not like sending water through a pipe, where water goes in one end and comes out the other. For instance, one cannot just “send” phonological representations to semantics. Phonological representations per se are completely unintelligible to semantics; and retinal images per se mean nothing to “higher-level” visual representations such as those involved in face recognition. Hence, even if one notates the interface with an arrow, it has to signify a complex correspondence or conversion between one sort of data structure and another, and this conversion is an essential component of mental computation.

In order to describe cognition, then, it is important to characterize both the data structures in the head and the interfaces among them. At least for the foreseeable future, we can do so only through reverse engineering, through developing theories of mental representations that enable us to understand subjects’ behavior (including their reports of their experience—Dennett’s (1991) heterophenomenology—which is also a kind of behavior).

In short, a theory of mental computation has to characterize the repertoire of domains (or “levels”) of mental representation, such as phonological structure and whatever kinds of structures are involved in encoding meaning. This repertoire may be different for different species. For instance, we probably should not expect kangaroos to have phonological structure in their repertoire. But we (or at least I) would suspect that all humans share a common repertoire of domains of mental representation.

This is not to deny that individuals may differ significantly in the richness of their representations in particular domains. A professional soccer player must have much more highly articulated motor representations than I do, but I certainly don’t lack motor representations altogether. In fact, I probably have myself developed some highly specialized motor representations connected with playing the clarinet.<sup>2</sup>

There need to be (at least) two kinds of computations over mental representations. One kind remains within a domain: computing a new representation in the domain on the basis of existing ones. For example, the phonology of a suffix can be adjusted, depending on what it is attached to: the plural of *cat* is /kæts/, with an /s/, but the plural of *dog* is /dɒgz/, with a /z/. Similarly, mental rotation involves computing new visual representations on the basis of existing ones.

---

2. Are individuals with autism entirely lacking in whatever representations are responsible for theory of mind, or are their representations just sparser?

The other kind of computation uses the interface links between two domains  $D_1$  and  $D_2$  to compute a structure in  $D_2$  on the basis of a structure in  $D_1$ . This is the sort of computation involved in any sort of perception, for instance using a phonological input to compute a sentence meaning from the meanings of the individual words and their phonological linear order. Speaking requires a mapping in the opposite direction; reading uses visual representations to compute linguistic structures (primarily phonological), via interface links between the two domains.

Hand in hand with identifying different domains of mental representation and their interfaces, another issue must be addressed: What determines the potential repertoire of data structures *within* each of the domains? And if there is an interface between two domains, what determines the potential repertoire of links? If a domain is finite, the repertoire of structures can be enumerated. But in most any domain of interest, such as vision, language, music, social cognition, or action, an enumeration is impossible. The repertoire is potentially open-ended: the organism is capable of creating novel data structures. Yet such new structures are not unconstrained; they fall into patterns. So an essential issue for cognitive theory is to characterize these patterns. In the domains involved in the language faculty, these amount to what Chomsky (1986) calls “knowledge of language.” The issue I will be concerned with here is how these patterns are to be characterized formally.

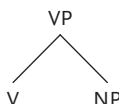
But first, just a little more context is in order. First, as Chomsky (1965) stresses, there is a further issue of how the patterns that constitute “knowledge of language” get into the brain, that is, how knowledge of a language is *acquired*. Recurring disputes within linguistics, as well as between linguistics and much of cognitive science, concern the cognitive resources necessary for acquiring language—so-called Universal Grammar—and how *they* are to be characterized. Are some of these resources specific to the language faculty, or are they all domain-general? I won’t rehearse the debate here (but see Jackendoff, 2002, chapter 4). In any event, whatever these resources are, whether language-specific or domain-general, they have to be wired into the brain prior to learning, since they form the basis for learning.

A further and hugely difficult issue is how all these mental representations and computations over them are instantiated in neural tissue—what Marr (1982) calls the “implementational level.” This is more than just a matter of finding brain localization; ideally it involves showing exactly how neurons encode, store, and compute these data structures and their links. And beyond this, there is the issue for developmental biology of how a developing brain comes to have the potential for computation that it does, under genetic guidance and in response to environmental input.

## 2. Three Models of Linguistic Rules

How are the patterns of mental representations for language to be characterized formally? I want to contrast three major approaches, which I'll call the Procedural Strategy, the Declarative Strategy, and the No Rules Strategy. This section lays them out; the rest of the paper is devoted to deciding among them.

The Procedural Strategy (also called “proof-theoretic” or “generative enumerative,” Pullum, 2013) defines the potential repertoire of linguistic representations by means of *derivations* that build linguistic expressions step by step. This is the approach of classical generative grammar and all its descendants in the narrow Chomskyan tradition, including the Minimalist Program (Chomsky 1995, 2002) and Distributed Morphology (Halle & Marantz, 1993; Siddiqi forthcoming) as prominent contemporary subtraditions. In this approach, rules of grammar are taken to be procedures that are available for building structure. For instance, the phrase structure rule (1a) is an instruction to rewrite a VP symbol as a sequence V—NP, or to expand VP into the tree (1b).

- (1) a.  $VP \rightarrow V - NP$   
 b. 

Similarly, the rule determining the form of English plurals can be stated informally as a directive: “To form the plural of a noun, add /z/.” The open-endedness of linguistic representations comes from the fact that some rules of grammar are recursive, that is, the rewriting of some symbols such as S and NP can eventually result in an output containing another instance of the same symbol.

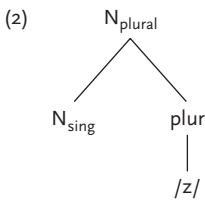
Since the early days of generative grammar, it has been recognized that these procedures cannot possibly correspond to the way language is actually processed in the brain. A speaker surely does not create a sentence by successively expanding S into a syntactic tree,<sup>3</sup> then “sending” this (a) to semantics to determine what it means and (b) to phonology to determine how to say it. Standard generative grammar therefore adopts the stance that the linguist’s grammar should be thought of not as an active procedure, but simply as a convenient way of defining a set, parallel to inductive definitions of number in mathematics. In turn, this stance leads to the familiar firewall between competence (the characterization of linguistic patterns) and performance (the actual processes taking place in the brain), going back to Chomsky (1965).

---

3. Or, in the Minimalist Program, starting with a collection of words and building upward to S; or, in Distributed Morphology, deriving the completed syntactic tree first and then populating it with words.

The Declarative Strategy (some versions of which are called “model-theoretic” by Pullum, 2013) defines the repertoire of possible linguistic representations in terms of a set of *well-formedness conditions* on structure. This approach encompasses frameworks such as Lexical-Functional Grammar (LFG; Bresnan, 2001), Head-Driven Phrase Structure Grammar (HPSG; Pollard & Sag, 1994), Construction Grammar (Goldberg, 1995; Hoffman & Trousdale, 2013), Autolexical Syntax (Sadock, 1991), and the Parallel Architecture (Culicover & Jackendoff, 2005; Jackendoff, 2002), among others.

On this approach, a phrase or sentence as a whole is legitimate if all its parts are *licensed* by one of the well-formedness conditions. For example, the rule for the English transitive verb phrase is not a rewriting rule, but simply the piece of syntactic structure (1b) itself: a syntactic schema or “treelet.” Any part of a syntactic structure that conforms to this structure is thereby licensed by the rule. Similarly, the English plural rule is stated as a *schema* that licenses plural nouns, along the lines of (2).<sup>4</sup>



Here, the open-endedness of linguistic forms is a consequence of what might be called “structural recursion”: a complex syntactic (or semantic) structure, licensed by a particular schema, can contain subparts that are licensed by the very same schema.

Well-formedness conditions can also implement correspondences between two domains. For instance, one such condition stipulates that an agent of an action (a semantic configuration) is typically expressed by the subject of the verb that denotes the action (a syntactic configuration).

Unlike the Procedural Strategy, the Declarative Strategy ascribes no implicit directionality to linguistic representations. There is no “input” or “output”; nothing is “sent” to another component for further steps of derivation. For instance, it is impossible to stipulate that words are “inserted” into sentences “before” or “after” some step of derivation.

4. I am abstracting away here from the variation between /z/, /s/, and /əz/. In a procedural approach, the plural has an “underlying form,” say /z/, and the grammar has rules that in the course of derivation change this to /s/ or /əz/ in the appropriate contexts. In a declarative framework, the plural morpheme simply has three forms that are licensed in different contexts.



Paradoxically, this makes a declarative grammar potentially more amenable to being embedded in a model of language processing. In processing, the grammar is being used to build up a linguistic structure online, over some time course (Jackendoff, 2002, chapter 7; Jackendoff, 2007). A procedural grammar determines an order in which rules are to be applied; and as seen above, this order makes little sense in terms of actual sentence processing. By contrast, in a declarative grammar, the processor builds structure by “clipping together” stored pieces of structure in any order that is convenient, using a procedure called unification (Shieber, 1986), which is arguably domain-general (Jackendoff, 2011).<sup>5</sup> In particular, in sentence comprehension, phonological structure can be taken as “input” and semantics as “output”; and in sentence production, the reverse. A declarative grammar doesn’t care in which direction it is put to use.

The No Rules Strategy encompasses a variety of theories, including connectionism (Christiansen & Chater, 1999; Elman, 1990; Rumelhart & McClelland, 1986; Seidenberg & MacDonald, 1999), item- or exemplar-based theories (Bod, 2006; Frank, Bod, & Christiansen, 2012; Pierrehumbert, 2002;), and statistically based approaches such as Latent Semantic Analysis (Landauer & Dumais, 1997). What they have in common is a rejection of the idea that the repertoire of linguistic forms is determined by any sort of abstract schemas such as phrase structure rules. Rather, knowledge of language consists of a list or an amalgam of forms one has encountered, and the open-endedness of language is ascribed to processes that construct or interpret novel expressions by analogy to stored forms.

This approach faces an inherent logical problem. Even if one’s knowledge of language lacks rules or schemas, one still has to somehow be able to encode the potentially unlimited variety of instances one encounters and remembers. So the basic question arises: In what format are instances encoded? In order to store and compare tokens, they must be coded in terms of a set of features and/or a set of analog dimensions of variation. The basic tenet of the No Rules Strategy is that such a repertoire cannot be explicit, because that would amount to having rules that define possible instances. So the characterization of the repertoire has to be implicit, built into the architecture in such a way that the distinct domains of mental structure emerge from the input.<sup>6</sup> The crucial question is how

---

5. Unification is a sort of Boolean union defined over hierarchical structures rather than sets. For instance, the unification of the string ABC with the string BCD is the string ABCD, overlapping at BC—rather than, say [ABC][BCD] or some other combination.

6. Just saying the dimensions of variation are “emergent properties” is not enough. For instance, some aspects of phonology are motivated by (and hence “emerge from”) acoustic and motor affordances. But they are not inevitable. German and Dutch have a rule of “final devoicing,” such that some words end in an unvoiced consonant at the end of a word, but use the corresponding voiced consonant when followed by a vowel suffix; hence German *Hand*, “hand,” is

phonological tokens (for example) come to have such different dimensions of similarity from, say, semantic tokens.

In the Procedural and Declarative Strategies, the variables in rules are the essential source of open-endedness. VP, for instance, stands for *any* VP, and N<sub>plur</sub> stands for *any* plural noun. The No Rules Strategy will have none of this. It insists that knowledge of language involves no abstraction of the sort captured by variables. Unfortunately, Marcus's (1998, 2001) critique of connectionist approaches shows that a network of the usual sort is in principle incapable of learning and encoding the sorts of relations usually taken to involve two variables—relations as simple as “X is identical to Y”—in such a way that they can be extended to novel instances outside the boundaries of the training set. Yet two-place relations are endemic to linguistic structure: “X rhymes with Y,” “X agrees with Y in number,” “X is coreferential with Y,” and so on. Hence there are strong reasons to believe that connectionist machinery is not up to the task of learning language in general. (And I believe similar arguments apply to other varieties of the No Rules Strategy. See also Jackendoff, 2002, section 3.5.)

The No Rules Strategy has been most successful in dealing with the structure of individual words, starting with Rumelhart and McClelland's (1986) well-known treatment of the English past tense, and in dealing with the impressive degree of frequency- and contextually-induced micro-variation in phonetic realization (as documented by Bybee, 2010) and lexical access (as studied by Baayen, 2007). However, this strategy has been explored and tested almost exclusively on relatively simple morphological phenomena such as English past tense formation and on the syntax of toy languages with minimal structure and a minimal vocabulary. To my knowledge, the approach has not been scaled up to full complex sentences such as this one, whose structure is relatively well understood in terms of either procedural rules or declarative schemas. The basic difficulty seems to be that a No Rules Strategy offers no way to encode and build hierarchical structure, a fatal flaw in a theory of language.<sup>7</sup>

---

pronounced /hant/, but its plural *Hände* is pronounced /hendə/, with phonetic [d] rather than [t]. This alternation can be motivated on physiological grounds: devoicing is a natural outcome of lowered air pressure at the end of a stressed syllable. One might therefore want to argue that there is no need for phonology—that apparently phonological principles are “emergent” from physiology. However, final devoicing is still a phonological principle—part of the grammar of German and Dutch—not a purely physiological one. If it were purely physiological, English would have to do it too, and it doesn't: *hand* is pronounced with a voiced [d]. So the physiology motivates but does not determine phonology (Blevins, 2004).

7. Frank, Bod, and Christiansen (2012) attempt to argue that language lacks hierarchical syntactic structure, but they build hierarchy implicitly into their parsing procedure, and they deal with only the simplest nonrecursive sentence structures.

I have tried in this section to cite relatively pure examples of each strategy. But there are hybrids as well. Tree Adjoining Grammar (Joshi, 1987) is declarative, in the sense that its building blocks are pieces of structure. But the combination of these pieces is still conceived of procedurally. Optimality Theory (Prince & Smolensky, 2004) is based on well-formedness constraints rather than rewriting rules, so in that sense it is declarative, but it still has “input” and “output”—procedural notions—as essential theoretical constructs.

### 3. Words and Rules Belong in the Same Bucket

How to adjudicate among these three approaches? This section offers an argument for the Declarative Strategy and against the Procedural Strategy; we return to the No Rules Strategy in subsequent sections. The argument “turns the knobs” on the distinction between words and rules, spotlighting linguistic phenomena that fall in the cracks between the two.

A typical dictionary of a foreign language has several hundred pages of words plus a small separate section devoted to grammar. From the very beginning, generative grammar inherited this traditional distinction, assuming (or even asserting—Chomsky, 1965) that the lexicon and the grammar are fundamentally different kinds of mental representations: words are declarative and rules of grammar are procedural. I want to argue against this assumption and show that in giving it up, there is nothing to lose (aside from tradition), and there is actually a great deal to gain.

The argument goes by a slippery slope: there are things that have to be stored in the lexicon, in declarative format, that are progressively more and more rule-like, so there seems less and less reason to distinguish them from things that everyone accepts as rules. Not only is there no clear place to draw the line between words and rules, there is no need to: when you get to the bottom of the slippery slope, you discover it’s not so bad down there after all. (Versions of this argument have been made by HPSG, Cognitive Grammar (Langacker, 1987), and Construction Grammar, as well as by me, in Jackendoff, 2002).

Starting with words: In any mentalist theory, words are stored in memory, presumably in some declarative form: the phonological structure /kæt/ is linked to some semantic structure CAT (which I leave uncharacterized here) and some syntactic features such as part of speech, as sketched in (3). (I don’t think anyone believes that the word *cat* is a procedure to derive CAT from /kæt/ or vice versa.)

- (3) Phonology: /kæt/  
 Syntax: +N, singular  
 Semantics: CAT

What about rules? The Procedural Strategy says they are procedures; the Declarative Strategy says they are pieces of structure that license parts of composed structures. But of course, the rules of the grammar, like the words, have to be present in memory in some form or another, either explicitly—or if only implicitly, then in the structure of the processor. Either way, we might want to say the lexicon and the grammar are stored differently in the brain. For instance, the lexicon might be stored in the temporal lobe and the grammar stored or otherwise instantiated in the frontal lobe, more or less as proposed by Ullman 2004.

But let us look at this distinction more closely. Example (3) is just the stereotypical case of a word. There are other variants. Words like those in (4) have phonology and semantics but no identifiable part of speech. They can occur alone as a full utterance, and they occur in combination only in things like *Hello, Bill* and in quotative contexts like (5), where anything at all can be inserted, even an expression in another language.

(4) *Phonology and semantics, no syntax:*

hello, ouch, yes, oops, dammit, upsey-daisy, allakazam, feh, uh-oh, shucks

(5) “Hello,” she said. (cf. “*Shema Yisroel*,” she said.)

English also has a few words such as those in (6) that have phonological and syntactic features but no semantic features. They function just as “grammatical glue.”

(6) *Phonology and syntax, no semantics:*

- |                                   |                              |
|-----------------------------------|------------------------------|
| a. epenthetic <i>it</i> :         | <b>It</b> 's noisy in here.  |
| b. <i>do</i> -support <i>do</i> : | <b>I didn't</b> see her.     |
| c. <i>N of NP</i> :               | a picture <b>of</b> Bill     |
| d. subordinating <i>that</i> :    | I know <b>that</b> you came. |

Nonsense words like those in (7) have phonology, but no syntax or semantics at all; their linguistic function is just to fill up metrical space in poetry or songs. If you know the song, you recognize these the same way you recognize words.

(7) *Phonology, no syntax or semantics*

fa-la-la, hey diddle diddle, e-i-e-i-o, brrr-raka-taka, inka-dinka-doo, rickety-tickety-tin

So we find all different combinations of phonological, syntactic, and semantic features in stored lexical items.

What else has to be stored in the lexicon? It is clearly necessary to store *idioms* and other fixed expressions in some form or another. For instance, *kick the bucket* has phonological structure, plus a semantic structure approximately equivalent to *die*, plus the syntactic structure of a verb phrase. These have to be stored as a unit because the meaning is noncompositional.

- (8) Phonology: /kɪk#ðə#bʌkət/  
 Syntax: [<sub>VP</sub> V [<sub>NP</sub> Det N]]  
 Semantics: DIE (X)

We know that (8) has the structure of a verb phrase, and is not simply an undigested string, because *kick* inflects just like an ordinary verb, e.g., *he kicked the bucket*, not *\*he kick-the-bucketed*.

Another reason to think idioms have internal syntactic structure is that, just like verbs, they can have argument structure. For instance, the idioms in (9) take a freely chosen direct object, just like ordinary transitive verbs, and this direct object goes exactly where a direct object *should* go—which happens to be in the middle of the idiom.

- (9) take NP for granted  
 put NP on ice  
 give NP the once-over

A theory of knowledge of language cannot shrug idioms off as some kind of exceptional aberration. English has thousands of idioms, perhaps tens of thousands. Hence it is necessary to give up the assumption that declarative memory for language—the lexicon—is simply a list of words: it also contains items with internal syntactic structure.

Other phenomena use standard syntax, but to unusual semantic ends. Consider (10).

- (10) Willy whistled his way down the hall.

Syntactically, this sentence has a typical verb phrase. However:

- The normal direct object of the verb *whistle* denotes an acoustic phenomenon (*whistle a tune* or *a signal*), but instead, (10) has this strange phrase *his way* in object position.

- *Whistle* also doesn't usually occur with a path expression like *down the hall*. (\**Willy whistled down the hall*.)
- The meaning of the sentence involves Willy *going* down the hall, even though there is no verb of motion.
- What *whistle* is doing is describing the manner or concurrent activity with which he goes down the hall.

In other words, (10) has a non-canonical mapping between syntax and semantics, marked by the phrase *his way* in object position. Any verb of the right semantic type is possible: you can drink your way across the country or knit your way through a conference. This sort of phenomenon might be called a “constructional idiom,” and the case in (10) might be sketched informally as (11).

(11) *Way*-construction:

Syntax/Phonology: [<sub>VP</sub> V *pro*'s way PP]

Semantics: 'NP goes PP while/by V-ing'

(12) gives two other mappings of this sort, with different semantics, marked by *away* and *head* (or other body part) *off*.

(12) a. *Time-away* construction:

Syntax/Phonology: [<sub>VP</sub> V [<sub>NP</sub> (*time*)] away]

Semantics: 'NP spends/wastes NP V-ing'

e.g. *Fred drank the afternoon away*. (= 'Fred spent/wasted the afternoon drinking')

b. *Head off* construction:

Syntax/Phonology: [<sub>VP</sub> V *pro*'s head/tush/butt off]

Semantics: 'NP V-s a lot/intensely'

e.g. *Suzie sang her head off*. (= 'Suzie sang a lot/intensely')

Knowing these constructions is part of knowing English. For each one, a speaker has to learn and store something about its syntactic structure, something about how its constituents correspond to semantics in other than the normal way, and something about the phonology of the designated elements *way*, *away*, and *head off* that signal that something unusual is going on.

Other constructions of this sort, such as the four in (13), have no distinguishing phonological content, so they're not very wordlike at all.

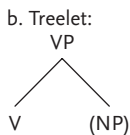
- (13) a. Sound+motion construction  
 [<sub>VP</sub> V PP]  
*The bus rumbled around the corner.* (= ‘The bus went around the corner, rumbling’)
- b. Inverted NP construction:  
 [<sub>NP</sub> a/this/that N of an N]  
*that travesty of a theory* (= ‘that theory, which is a travesty’)
- c. Light verb construction:  
 [<sub>VP</sub> V NP NP/PP]  
*Pat gave Kim a hug.* (= ‘Pat hugged Kim’)
- d. Casual paratactic conditional:  
 [S, S]  
*You break my chair; I break your arm.* (= ‘If you break my chair, I will break your arm’)  
 (Edward Merrin, June 12, 2014)

*The bus rumbled around the corner* means that the bus *went* around the corner—even though there is no verb of motion—and it was simultaneously making rumbling sounds. In *that travesty of a theory*, the syntactic head is *travesty*, and *theory* is a modifier; but semantically, *theory* is the referent of the expression and *travesty* is a modifier. In the light verb construction, the direct object *a hug* provides semantic content that is normally expressed by a verb, and the verb is more or less a dummy that provides argument positions and a host for tense. Finally, the paratactic conditional has no *if* or *then*, but the conditional meaning is perfectly clear.

Again, a speaker has to store knowledge of the constructional idioms in (13) as an association between a syntactic complex and a semantic complex. But the basic formalism is the same as for words—except that, as in idioms, the syntax is composite, not just a couple of features, and this time there is no phonology. The tradition of Construction Grammar (Goldberg, 1995; Hoffmann & Trousdale, 2013) has studied a large number of constructions of the type in (12) and (13), as have I (Culicover & Jackendoff, 2005; Jackendoff, 2010). Languages are full of them.

We now reach the crucial point in the argument. Using the very same formalism, we can state a language’s phrase structure rules as syntactic schemas, without any association to phonology or semantics. For example, (14) is a partial rule for the transitive verb phrase in English. Tomasello (2003) calls this sort of thing a “slot-and-frame schema,” using the notation in (14a); Janet Fodor (1998) calls it a “treelet,” in the alternative notation (14b), which is our original structure for the VP in (1b).

- (14) a. Slot-and-frame schema:  
 [<sub>VP</sub> V – (NP) ...]



In light of what we have seen up to this point, (14) should be thought of as another type of lexical item. There are lexical items with just phonology like *fid-dlededee*, items with just phonology and syntax like the *do* of *do*-support, items with just phonology and semantics like *hello*, and items such as idioms and meaningful constructions with complex syntactic structure. (14) happens to be an item with only syntactic structure—just one more logical possibility in the system. Its variables license novel, freely combined VPs in syntax.

At this point we have slid to the bottom of the slippery slope from words to rules, and it actually makes some things work better. The lexicon—the collection of declarative knowledge of language—includes both words like *cat* (3) and phrase structure rules like that for VP (14), as well as all sorts of things in between, such as idioms and meaningful constructions. The standard approach of generative grammar, in which words are declarative knowledge and rules are procedural knowledge, has to draw an unnatural line somewhere among these intermediate cases (and in practice, it has ignored them). In contrast, in a declarative conception of rules of grammar, this continuum is perfectly natural.

This leads to the radical hypothesis that all phenomena that have previously been called rules of grammar can be reformulated as pieces of structure stored in the lexicon. In other words, knowledge of language is wholly declarative, and the only procedural component—the only “operation on mental representations” that assembles larger structures—is unification. Furthermore, unification can assemble structures within any domain of mental representation, so strictly speaking, it is not part of “knowledge of language” *per se*.

This hypothesis presents numerous challenges for linguistic theory. For one thing, the domains of morphology (word structure) and phonology (sound structure) have nearly always been worked out in procedural terms, and it remains to be demonstrated that they can be reformulated in terms of purely declarative rules (see Jackendoff & Audring, 2016, forthcoming, for a beginning). Moreover, within syntax itself, the procedural notions of movement and deletion have to be expunged from the syntactic armamentarium. So it is crucial to figure out how a declarative theory of rules deals with phenomena such as passive, subject-auxiliary inversion, *wh*-question formation, and so on, which have always seemed very natural under a movement hypothesis. There is no space here to go through the arguments, but HPSG, LFG, Simpler Syntax (Culicover & Jackendoff, 2005), and various other frameworks are all attempts to work this out, with a good deal of success.

#### 4. Full and Partial Productivity

Twirling a different knob on the lay cognitive scientist’s conception of language, we next consider rules governing *morphology*, the interior structure of words. Traditional grammar tends to view morphology procedurally. For instance, it is



natural to think of deriving the word *procedurally* by adding the suffix *-ly* to the word *procedural*, which in turn is derived by adding the suffix *-al* to the word *procedure*, which is itself derived by adding *-ure* to *proceed*. The earliest generative treatments of morphology, such as Lees (1960), took over this assumption, deriving morphological structure by means of transformations. For instance, Lees had transformations that turned the sentence *John constructed a theory* into the NP *John's construction of a theory*.

The difficulty with this approach is that transformations are supposed to be completely regular and productive, but lots of morphology *isn't* regular or productive. Three symptoms appear over and over again.

Symptom 1: In many cases, although evidently a rule is involved, the instances of the rule still have to be listed individually. For example, many “derived” nouns consist of a verb plus the suffix *-ion*, and many more consist of a verb plus *-al*. But these suffixes cannot be applied indiscriminately, in the fashion that standard rules would dictate: *confusion* and *refusal* are real words, but the perfectly plausible *\*refusion* and *\*confusal* are not. These idiosyncratic facts must be recorded somewhere in the grammar or the lexicon.

Symptom 2: Stereotypically, so-called “derived” items have words as their roots: *construction* is based on *construct* and *permission* is based on *permit*. But there happen to be many “derived” items without a lexical word as their “root,” such as *commotion*, *tradition*, *ambition*, and *retribution*. This is not a matter of rare exceptions. For instance, among the hundreds of English adjectives that end in *-ous*, over a third of them lack a lexical root, for example *tremendous*, *gorgeous*, *salacious*, *impetuous*, and *supercilious*.

Symptom 3: The meanings of so-called “derived” items are often partly idiosyncratic with respect to the meaning of their roots, so they cannot be generated by a general meaning-blind procedure. Consider for instance the difference between the meanings of *recite* and *recital*, or *proverb* and *proverbial*. Again, this phenomenon is extremely widespread.<sup>8</sup>

I will use the term “partially productive” for patterns with some combination of these three symptoms. Such rules contrast with fully productive patterns, which

---

8. A common impulse is to appeal to the history of English, for example, borrowings from Latin, to account for these facts. However, speakers of English are typically unaware of the origin of the words they learn. So on the whole, the history of the language plays no role in its speakers' knowledge.

apply to *every* eligible form aside from listed exceptions, and whose meanings are completely predictable. Speakers are happy to apply fully productive patterns to novel items, for example unhesitatingly giving the plural of *wug* as *wugs* (Berko, 1958).

Chomsky and Halle (1968) and Lakoff (1970) retained the Procedural Strategy, and accounted for partially productive phenomena by means of “exception features.” For instance, *refuse* would be marked as unable to undergo the *-ion* transformation, and *tremend* would be marked as obligatorily undergoing the *-ous* transformation. While this approach deals with symptoms 1 and 2 by brute force,<sup>9</sup> it does not address symptom 3.

Chomsky’s (1970) solution to partial productivity, the so-called Lexicalist Hypothesis, was to put partially productive morphology “in the lexicon,” prior to inserting words into sentence structures. He was not very specific about how this would work; but at the time, people thought about his proposal in terms of “lexical rules”—rules in the lexicon distinct from rules of phrasal grammar. For instance, Wasow (1977) produced widely accepted arguments that there is a distinction between fully productive syntactic passives such as (15a), produced by syntactic rules, and partially productive adjectival passives such as (15b), produced by lexical rules.

- (15) a. Pat was hugged by Kim.  
b. Pat was very surprised at the news.

Some subsequent syntactic frameworks such as Lexical Functional Grammar (Bresnan, 1982, 2001) and Head-Driven Phrase Structure (Pollard & Sag, 1987) have taken this distinction as foundational, as have some morphological theories (e.g., Anderson, 1992, Stump, 1990).

Some approaches to morphological rules (e.g., Aronoff, 1976; Halle, 1973; Halle & Marantz, 1993; Lieber, 1992) take them to be encoded as procedural rules that create morphologically complex words such as *procedurally* from simpler words or stems stored in the lexicon. However, if one takes symptoms 1–3 seriously, a procedural account of partially productive rules is problematic in the same way as Lakoff’s account. First, it still has to somehow list which affixes can be added onto *refuse* and which ones onto *confuse*. Second, a word like *tremendous* has no root in the lexicon to which the suffix *-ous* can be added, so it is

---

9. Actually, to a first approximation only. For example, *ambi* would have to be marked to obligatorily undergo *either* the *-tion* transformation (to form *ambition*) or the *-ous* transformation (to form *ambitious*). In other words, one would have to invoke Boolean combinations of exception features.

unclear how to generate it (unless *\*tremend* is listed as requiring a suffix). Finally, putting *recite* and *-al* together does not alone specify what *recital* means.<sup>10</sup>

Other approaches (e.g., Blevins, 2006; Bochner, 1993; Booij, 2010; Jackendoff, 1975) take lexical rules to be declarative patterns that characterize relations among full words stored in the lexicon. For instance, both *refuse* and *refusal* are stored in the lexicon, and the *-al* rule does not *derive* one from the other, but rather stipulates how these forms are related. The *-al* rule itself is stated as a schema containing variables: it expresses the possibility for a noun to consist of a verb plus the suffix *-al*, as in (16). (*V* in syntax is a variable for the verb, *X* is a variable for its meaning in semantics, and / . . . / is a variable for its phonological form.)

- (16) Phonology:     / . . . / – al  
       Syntax:       [<sub>N</sub> V – suffix]  
       Semantics:    ACT-OF X-ING *or* RESULT-OF X-ING

This approach fares somewhat better with the three symptoms of partial productivity.

Symptom 1: Since the declarative theory lists all the forms that exist, the fact that there is no word *\*refusion* has nothing to do with features of *refuse*; there simply is no such word listed in the lexicon. To be sure, the lexical rule for *-ion* says that if there *were* such a word, it could be related to *refuse*; but it does not predict that there *is* such a word.

Symptom 2: The fact that there *is* a word *tremendous* but no *\*tremend* is the mirror image of this situation: the *-ous* rule says that if there were such a word as *\*tremend*, it could be related to *tremendous*. But again, it does not *require* there to be such a word.

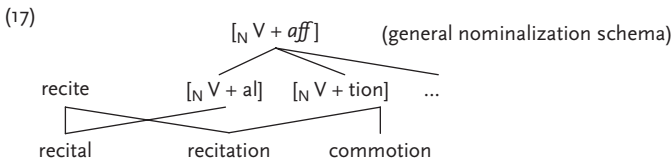
---

10. Some strands of generative grammar, such as Hale and Keyser (1993, 2002), revert to accounting for morphologically complex items through syntactic operations such as head-raising. For instance, they derive *saddle the horse* from something like *put a saddle on the horse*. But their derivational mechanism has the very same three problems as Lees's approach: how is application of the rules restricted to only certain items, how is something derived when there is no root to derive it from, and how are semantic idiosyncrasies accounted for? To my knowledge, none of these three problems has been seriously addressed within the Hale & Keyser approach.

Distributed Morphology (Halle & Marantz, 1993; Harley 2014; Siddiqi forthcoming) accounts for symptoms 1–3 by appealing to a Vocabulary list of existing forms and an Encyclopedia list of word meanings—more or less the position taken here, but distributed across a procedural derivation. However, I am not aware of any formal characterization of the Vocabulary or the Encyclopedia within this framework.

Symptom 3: The fact that *recital* means something other than ‘act of reciting’ is tolerated by the lexical rule, which motivates but does not fully determine the meaning relation between the two words. In other words, lexical rules can be considerably more flexible if stated declaratively than if stated procedurally.

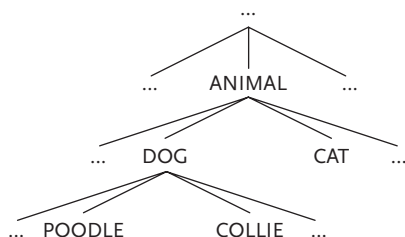
How does a morphological schema relate forms such as *recite* and *recital*? An approach that has been broadly adopted (HPSG: Pollard & Sag, 1994; Construction Grammar: Goldberg, 1995, 2006; Construction Morphology: Booij 2010; Parallel Architecture: Jackendoff, 2002) treats words and morphological schemas as nodes in an inheritance hierarchy. A word “inherits” structure from smaller words and more abstract schemas that predict parts of its form. The predicted parts of its structure are redundant, and therefore in some sense “cost less”: there is less to learn.<sup>11</sup> For instance, as shown in (17), *recitation* inherits structure both from the word *recite* and from the deverbal *-tion* schema—but it also has idiosyncratic semantic structure of its own which must be “paid for.” An item such as *commotion*, which lacks a lexical root, also inherits structure from the *-tion* schema. But since there is no root *commote*, it has to “pay” for its pseudo-root. As shown by Bochner (1993), schemas can also inherit structure from more abstract schemas. For instance, the *-tion* schema inherits structure from the more general nominalization schema, which is also ancestor to other nominal schemas such as *-al*.



One thing that makes the notion of inheritance hierarchies attractive is that the same sort of machinery is needed for the relations among concepts: the concept of *animal* is ancestor of concepts like *cat* and *dog*; *dog* is the ancestor of *poodle* and *collie*, and so on, as in (18) (Murphy, 2002). In other words, this sort of taxonomic organization stands a chance of being domain-general, hence not specifically part of “knowledge of language.”

11. The notion of “cost” is played out differently in different theories. See Bochner, 1993; Jackendoff 1975; Jackendoff and Audring (forthcoming).

(18)



Inheritance hierarchies raise many tricky questions—both formally and psycholinguistically (Jackendoff & Audring, 2016, forthcoming). Setting these issues aside, what is interesting is that they allow us to formalize a morphological rule like the *-al* rule as a piece of linguistic structure containing variables, as in (16). There is no need for a new kind of formal rule: like the idioms and constructions discussed in section 4, these schemas are lexical entries too.

However, we face a new problem: the very same format can be used for *fully* productive morphology. For instance, English present participles, which are completely regular, can be defined by the schema (19) (the meaning is assigned by other principles).

- (19) Phonology: / ... / - ing  
 Syntax: [<sub>Vprespart</sub> V - suffix]

For this sort of fully productive morphology, the standard assumption is that *not* every form is stored as a descendant in the inheritance hierarchy. When speakers encounter a new verb like *skype*, they know automatically that its present participle is *skyping*, and someone hearing this for the first time doesn't bat an eyelash. So this case looks more like standard syntax, and it should be “in the grammar,” not “in the lexicon”—just like the schema [<sub>VP</sub> *V-NP*]. But if partially productive lexical rules have the same format as fully productive syntactic rules, how can we tell them apart? Should the fully productive rules be “in the grammar” and the partially productive rules be “in the lexicon,” as Chomsky posited?

A radical answer, proposed by the connectionists (e.g., Bybee & McClelland, 2005; Rumelhart & McClelland, 1986), is that there are *no* explicit rules or schemas for morphological patterns—either partially or fully productive; there are only statistical tendencies based on the amalgamation of learned examples. As we saw in section 3, the larger claim of the No Rules Strategy is that no rules are necessary for *any* patterns in language: memory encodes only individual words (at best), and novel inputs are understood simply by analogy with or by interpolation among memorized instances.

The psycholinguists, led by Steven Pinker (e.g., Pinker, 1999; Pinker & Prince, 1988), dispute the No Rules position on psycholinguistic grounds that we need

not go into here, as well as on theoretical grounds, some of which were mentioned in section 3. They argue instead for a hybrid “dual-route” model, in which rules are indeed necessary for fully productive patterns like *-ing* words. However, they grant the connectionists’ position for partially productive patterns like *-al* words: these have no rules, only statistical patterns of instances.<sup>12</sup>

## 5. Dual Route—But Where?

So we are now faced with a three-way choice between Procedural, Declarative, and No Rules strategies, in two distinct domains: fully productive and partially productive patterns. Table 1 sums up the situation. Three of the six possibilities have already been ruled out by considerations raised in sections 3 and 4.

Some theories are homogeneous or uniform, in that they use the same types of rule for fully and partially productive patterns. For instance, Lees’s (1960) theory is homogeneously procedural; the connectionist theory is homogeneously No Rules (in fact, it denies that there is any distinction between fully and partially productive patterns).

Other theories are hybrid, using different types of rules for the two. For instance, in Chomsky’s Lexicalist Hypothesis, fully productive rules are in the grammar and are procedural, while partially productive rules are in the lexicon and ambiguously procedural or declarative. Pinker’s dual-route theory is another hybrid. It takes the fully productive rules to be procedural: “To form the past tense of a verb, add *-d*”; but there are no rules for partially productive patterns. Many contemporary morphologists, for example, Spencer (2013), adopt a similar hybrid position.

Jackendoff (2002) essentially adopts Pinker’s hybrid position, but substitutes declarative rules like (19) for Pinker’s fully productive procedural rules. I have

**Table 1. Possible positions on the nature of rules**

	Fully productive patterns	Partially productive patterns
Procedural	Not possible (shown in section 4)	Not possible (shown in section 5)
Declarative	Possible (shown in section 4)	Possible
No Rules	Not possible (shown in section 4)	Possible

12. I reluctantly fault Pinker for choosing to fight the connectionists on their own ground, without forcing them to face up to the larger challenges of hierarchical structure in phrasal syntax. Even if one can eliminate rules for English past tenses, this does not show that they can be eliminated for verb phrases and relative clauses. Pinker could have invoked these considerations in his argument, opening the debate immediately to a broader domain.

since become skeptical of this position, because it entails that, for instance, there is no *-al* suffix in the grammar of English, only a lot of *-al* words. This now strikes me as too deflationary. I would like to be able to say that knowledge of English includes an explicit declarative *-al* schema along the lines of (16).

Of the remaining options in Table 1, the choice boils down to whether partially productive patterns are best accounted for in a declarative or a No Rules approach. If the former, there is a homogeneously declarative account of all rules of language.<sup>13</sup> If the latter, we retain Pinker's dual route account, but state fully productive phenomena declaratively instead of procedurally.

In order to decide between these two options, I need to work out the declarative option a bit further. So for the moment let us make the assumption that there *are* partially productive schemas like the *-al* rule (16), and ask how they could differ from fully productive schemas like the *-ing* rule (19).

Recall that Chomsky (1970) put the fully productive rules "in the grammar" and the partially productive rules "in the lexicon." But in the approach developed in section 4, fully productive rules are schemas in the lexicon, just like partially productive rules. So Chomsky's solution is not available to us. Moreover, since we are *assuming* that there are schemas for partially productive rules, Pinker's solution is not open to us either. What options *are* available?

One possibility would be to simply tag schemas with the features [*fully productive*] or [*partially productive*]. A partially productive schema such as the *-al* schema would require an enumeration of its instances in the lexicon. A fully productive schema such as the *-ing* schema would not have such a constraint: it would allow free application to novel forms that are not listed.

I wish to propose something a bit more delicate: the distinction between partial and full productivity is marked, not on entire schemas, but on the variables within schemas. I'll call a variable marked for full productivity an *open* variable, and one marked for partial productivity a *closed* variable, as illustrated in (20).

- (20) a. (Partially productive *-al* rule)  
     Phonology:       / ... /<sub>closed</sub> - al  
     Syntax:           [<sub>N</sub> V<sub>closed</sub> - suffix]  
     Semantics:       ACT-OF X<sub>closed</sub> -ING *or* RESULT-OF X<sub>closed</sub> -ING  
   b. (Fully productive *-ing* rule)  
     Syntax:           [<sub>Vprespart</sub> V<sub>open</sub> - suffix]  
     Phonology:       / ... /<sub>open</sub> - ing

13. Pending, of course, a demonstration that all morphology and phonology can be treated this way; see Jackendoff & Audring (forthcoming) for a start.

A schema with a closed variable captures generalizations among listed items, through the inheritance hierarchy. A schema with an open variable can be used to create forms on the fly, such as *imprinting* and *differentiating*—and even novel forms such as *wugs* and *skyping*. But in addition, a schema with an open variable can serve as an ancestor of stored items. For instance, the productive English plural schema is an ancestor of stored items like *scissors* and *trousers*. In other words, open variables can do what closed variables can do—*plus* create new forms. In a sense, one can think of fully productive rules as partially productive rules that have “gone viral.”

Turning the knobs again, here is one piece of evidence for this solution. English has four different principles for naming geographic features, illustrated in (21). Each one involves a name and a descriptive term (underlined) for the type of geographical feature.

- (21) a. Arrowhead Lake, Biscayne Bay, Loon Mountain, Claypit Pond,  
Wissahickon Creek  
b. Lake Michigan, Mount Everest, Cape Cod  
c. the Indian Ocean, the Black Sea, the Hudson River, the Ventura  
Freeway, the San Andreas Fault  
d. the Bay of Fundy, the Gulf of Aqaba, the Isle of Man, the Cape of  
Good Hope

Speakers of English know hundreds of place names of the sorts illustrated in (21). Their structure is inherited from the schemas in (22).

- (22) a. [<sub>N</sub> Name<sub>open</sub> N<sub>closed</sub>]  
b. [<sub>N</sub> N<sub>closed</sub> Name<sub>open</sub>]  
c. [<sub>NP</sub> the Name<sub>open</sub> N<sub>closed</sub>]  
d. [<sub>NP</sub> the N<sub>closed</sub> of Name<sub>open</sub>]

These schemas can be used freely to create new place names. If you want to name a mountain for Dan Dennett, you know immediately that you can call it *Dennett Mountain* (schema 22a) or *Mount Dan* (schema 22b). So the Name variable is completely open.

On the other hand, which of the four schemas has to be used depends on the word for the type of feature—whether it’s *lake* or *ocean* or *mountain* or *mount*. Speakers of English know which words go in which schemas: for instance, our mountain can’t be called *\*Daniel Mount* (schema 22b) or *\*the Mountain of Dennett* (schema 22d). So the variable for the type of geographic feature is only partially productive: the instances have to be learned one by one. Overall, then, these schemas have one variable of each type, and it is impossible to mark the schema as a whole as fully productive or not.



More generally, this case presents evidence against Chomsky's hypothesis that partial productivity is "in the lexicon" and full productivity is "in the syntax." In order to formalize this little subgrammar of place names, partial and full productivity must be in the same component of the grammar, as notated in structural schemas like (20) and (22).

I also conclude from this case that the difference between full and partial productivity—Pinker's dual route—need not be a matter of whether there is a rule or not, but can lie simply in the type of variable within schemas in the lexicon. In fact, one might say that there are indeed two routes: closed variables make use of one of these routes, and the open variables make use not just of the *other* route, but of *both*.

## 6. Aren't Morphology and Syntax Different?

It might be objected that this solution slides over an important difference between morphology and syntax: *morphology* may be full of these crazy partially productive patterns, but surely *syntax* isn't! If, as just proposed, the distinction between full and partial productivity in morphology is localized in the variables of schemas, then, since syntactic rules are also schemas, they could in principle be partially productive as well. At least according to the lay cognitive scientist—but also according to Chomsky's Lexicalist Hypothesis, this is wrongheaded, missing the point of what syntax is about.

One could resolve the issue by brute force: one could simply stipulate that variables in syntactic rules are all fully productive, but that those in morphological rules can be of either flavor. But this would not be the right resolution. Although syntax does tend to be more reliably productive than morphology, it has some strange little pockets that are genuinely partially productive. Consider again the schemas in (22c,d). We can tell these are patterns of phrasal syntax, because adjectives can be inserted in their usual position within NP: *the lordly Hudson River, the dangerous Bay of Fundy*. Yet, as just shown, one of the variables in each schema is closed.

Another such case is part of the English determiner system. There is no productive rule behind the list of the determiner patterns in (23).<sup>14</sup>

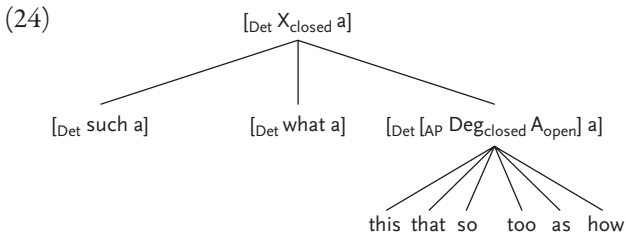
---

14. Plurals show further eccentricity:

- (i) a. such/what trees
- b. \*that beautiful trees (cf. *trees that beautiful*)
- c. \*how beautiful trees (cf. *\*trees how beautiful*)

- (23) a. such a tree  
 b. what a tree  
 c. that beautiful a tree  
 d. how tall a tree  
 e. \*very tall a tree  
 f. \*more beautiful a tree  
 g. \*twenty feet tall a tree

(24) shows the inheritance hierarchy for this construction.



The overall pattern is a determiner of the form  $X a$ , which attaches to a noun. This pattern is partially productive, and it has a small number of instances including *such a*, *what a*, and a sub-pattern containing an adjective phrase. Within the adjective phrase, the choice of degree word is partially productive, restricted to *this*, *that*, *so*, *as*, *too* and *how*. But the choice of adjective is fully productive: we can say things like *how tantalizing an idea* or *as ridiculous a movie as I've ever seen*.

In short, partial productivity extends across the entire lexicon, including syntax; it is not specific to morphology, as people often think.<sup>15</sup> By marking variables open or closed, it is simple to make the distinction between syntactically productive and partially productive patterns. In contrast, it is uncomfortable at best to generate patterns such as (22c,d) and (23) procedurally.

I conclude that a uniformly declarative theory can still maintain the distinction that Chomsky's Lexicalist Hypothesis and Pinker's dual route are meant to capture. But the distinction is not whether there is a lexicon-grammar distinction, as Chomsky proposes, or whether there are rules or not, as Pinker puts it. It is simply how the variables in the rule are marked for productivity.

15. For other cases, see Culicover (1999); Culicover and Jackendoff (2005, section 1.5); Kay (2013); Taylor (2012). Kay advocates for a version of Pinker's position: grammar concerns fully productive constructions only, while partially productive constructions belong instead to "meta-grammar," which operates by analogy rather than by rules.

## 7. Learning Partially and Fully Productive Schemas

But is this the right approach? Or can we get away with No Rules for partially productive patterns? To answer this, let us turn one last knob, and consider how a learner acquires schemas. They are not present in the input. At best, what *is* present in the input (once the learner has some command of phonology) is examples of utterances, which a schema-less learner must acquire and store in memory individually. The usage-based theorists (e.g., Bybee, 2010; Taylor, 2012; Tomasello, 2003) are perfectly correct on this point. However, the end state of acquisition has to include schemas for fully productive patterns like NP and VP. Where do they come from?

Various researchers from various schools of thought (e.g., Albright & Hayes, 2003; Culicover & Nowak, 2003; Tomasello, 2003) have conjectured that an active process in the learner's mind/brain looks for similarities among stored items. When this process finds similar items (or a sufficient number of similar items), it extracts the similar parts and builds them into a schema that replaces the differing parts with a variable. For example, if the lexicon contains items ABC, ABD, and ABE, this process adds a schema ABX to the lexicon, and ABC, ABD, and ABE listed as its daughters in the inheritance hierarchy.<sup>16</sup> The process can continue recursively, constructing more and more general schemas.

This approach, while appealing, raises a further question that has not to my knowledge been addressed clearly in the literature. Suppose language learners deduce a linguistic rule from the primary linguistic input. How do they decide whether a particular pattern is fully or only partially productive, and what is at stake?

In a hybrid theory of any sort, this choice entails deciding which component of the grammar this pattern belongs in—a rather radical choice. In Chomsky's approach, the choice is between being “in the syntax” and being “in the lexicon.” In Pinker's approach, the choice is whether the pattern is captured by a rule or whether there are merely associations. In contrast, the uniformly declarative approach proposed in section 6 makes the choice formally simple: the learner has to decide only which diacritic to place on a variable. That is, the choice is localized within a particular lexical entry—a much more straightforward distinction.

Culicover & Nowak (2003) and Tomasello (2003) propose (in our terms) that the learner initially treats a variable as closed, that is, that the default case is partial productivity. A partially productive schema would help learners understand and

---

16. There is the additional difficult problem of establishing the domain of the variable: Is it set to be maximally general, or to the minimum range needed to encompass the observed instances, or somewhere in between? See Albright and Hayes (2003).

store forms they have never heard before, but they would not be inclined to *produce* new forms without evidence for these forms in the primary linguistic data. This seems to accord with the lore about language acquisition: children, at least at first, are relatively conservative in their generalizations.

If the learner creates a schema and initially assumes it is partially productive, the next decision is whether to “promote” its variable from closed to open. Both kinds of variables can register patterns of relations among items in the lexicon, but a fully regular variable has two further capabilities. First, it can be used to create new forms such as *wugs* and *skypers* without prior evidence that they exist. Second, it eliminates the need to store new forms as they are heard. If one learns the word *patio*, one doesn’t need to also learn (or construct) the word *patios*; it can be understood the next time on the basis of its composition. So at the cost of using the non-default value of the closed-open feature, one gains the benefits of greater flexibility and less burden on long-term memory.

Now here is a crucial point of logic. The process that upgrades a variable from closed to open is not omniscient. It has no way of knowing in advance whether a pattern is going to be fully productive and or only partially productive. And it can’t wait until it has all the evidence: if it is considering setting up a schema, it has to store its hypothesis in order to test it against subsequent evidence.

What is the form of a hypothesis? A theory of mental representation requires us to ask this question—though it is rarely asked. In the present approach, the simplest way to encode a hypothesis is as a tentative schema. If the hypothesis is validated, the relevant variables in the schema are upgraded to open. But if it fails, that does not invalidate the original assessment that a pattern exists. Hence alongside the fully productive schemas, one should expect to find lots of partially productive schemas that have failed to be promoted. Since such schemas are capturing partial regularities, there is no reason to discard them. We should expect the lexicon to be littered with failed (but still not useless) hypotheses.

In short, partially productive schemas are potential steppingstones to fully productive rules. Fully productive schemas start life as partially productive schemas, and when they are promoted, they do not lose their original function of supporting inheritance hierarchies among listed items. Hence it stands to reason that there must be items with *just* the original function, namely partially productive schemas. This is altogether natural in a uniformly declarative theory of rules.

By contrast, a No Rules theory claims that there are no partially productive schemas that lead the way to the construction of regular rules. For hard-core No Rules theorists, this is not a problem, because in any event there are no regular rules to be constructed. But the failure of No Rules approaches to account for the free combinatoriality of language—which requires open variables—casts doubt on this option.

This conclusion also reinforces my doubts about Pinker's hybrid position, which grants the existence of rules for fully productive patterns but not for partially productive ones. If you need partially productive rules to attain fully productive rules, then you need partially productive rules, period.

In addition, this account of acquisition upends the notion, traditional within generative grammar and much of morphological theory, that only the fully productive rules are the real business of linguistic theory, and that the unruly "lexical rules" are relatively "peripheral." In the present approach, *all* patterns in language are expressed by schemas, of which a special subset happen to be fully productive. One cannot attain fully productive rules without a substrate of partially productive ones.

## 8. Ending

Returning to larger issues mentioned at the outset, the overall conclusion here is that:

- The mental representations involved in knowledge of language have the format of pieces of linguistic structure in phonology, semantics, and syntax (including both phrasal syntax and morphosyntax), plus linkages between them.
- What makes a piece of linguistic structure more rule-like is the presence of variables. Lexical items run the gamut from being fully specified (like *cat*) to being made up entirely of variables (like VP).
- Variables in either morphology or syntax can be marked as closed (partially productive) or as open (fully productive).
- The open-endedness of language is a consequence of combining pieces of structure that have open variables, using the domain-general operation of unification.

Looking to further horizons, this raises the question of whether domains of mental representation other than language can be characterized similarly. Candidate domains might include visual/spatial cognition, the formulation of action, and music. For instance, in music: How does one store one's knowledge of a melody? How is this knowledge supported by structural schemas at different levels of specificity, such as common metrical, melodic, and harmonic patterns? How does this knowledge allow one to appreciate ornamentation, elaboration, and variation on the melody? Among the structural schemas (if they exist), is there a counterpart of the fully/partially productive distinction? How are the structural schemas learned? And so on.

My working hypothesis is that the overall texture of different mental domains is largely the same, differing primarily in the types of units from which the domains are built and what other domains they connect with. Memory in all applicable domains is stored as pieces of structure—both fully specified items (like words) and underspecified items with variables (like rules and constructions). Open-endedness is achieved by unifying these pieces of structure. In this respect, the present argument, built on details of linguistic structure, presents a template and a challenge for cognitive science as a whole: Can this type of analysis serve to characterize all of cognition?

## Acknowledgments

Much of the material presented here has been developed in collaboration with Jenny Audring, my guide to the wilds of morphology and my co-author on many related publications. For the opportunity to try out many iterations of the arguments here, I am grateful to members of the Tufts linguistic seminar over the years: Ari Goldberg, Neil Cohn, Eva Wittenberg, Anita Peti-Stantić, Rob Truswell, Katya Pertsova, Anastasia Smirnova, Naomi Caselli, Rabia Ergin, and Geert Booij. Thanks also to Jay Keyser for many stimulating conversations on these issues. Finally, I owe a great debt of gratitude to Dan Dennett for an unending supply of intellectual adventures over the past 30 years.

## Works Cited

- Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, 90, 119–161.
- Anderson, S. (1992). *A-morphous morphology*. Cambridge, UK: Cambridge University Press.
- Aronoff, M. (1976). *Word formation in generative grammar*. Cambridge, MA: MIT Press.
- Baayen, R. H. (2007). Storage and computation in the mental lexicon. In G. Jarema & G. Libben (Eds.), *The mental lexicon: Core perspectives* (pp. 81–104). Amsterdam, Holland: Elsevier.
- Berko, J. (1958). The child's learning of English morphology. *Word*, 14, 150–177.
- Blevins, J. (2004). *Evolutionary phonology: The emergence of sound patterns*. Cambridge, UK: Cambridge University Press.
- Blevins, J. (2006). Word-based morphology. *Journal of Linguistics*, 42, 531–573.
- Bochner, H. (1993). *Simplicity in generative morphology*. Berlin, Germany: Mouton de Gruyter.
- Bod, R. (2006). Exemplar-based syntax: How to get productivity from examples. *The Linguistic Review*, Vol. 23, Special Issue on Exemplar-Based Models of Language

- Booij, G. (2010). *Construction morphology*. Oxford, UK: Oxford University Press.
- Bresnan, J. (Ed.). (1982). *The mental representation of grammatical relations*. Cambridge, MA: MIT Press.
- Bresnan, J. (2001). *Lexical-functional syntax*. Oxford, UK: Blackwell.
- Bybee, J. (2010). *Language, usage, and cognition*. Cambridge, England: Cambridge University Press.
- Bybee, J., & McClelland, J. (2005). Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *Linguistic Review*, 22, 381–410.
- Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
- Chomsky, N. (1970). Remarks on nominalization. In R. Jacobs & P. Rosenbaum (Eds.), *Readings in English transformational grammar* (pp. 184–221). Waltham, MA: Ginn. Also in Chomsky, *Studies on semantics in generative grammar*, The Hague: Mouton, 1972.
- Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. New York, NY: Praeger.
- Chomsky, N. (1995). *The minimalist program*. Cambridge, MA: MIT Press.
- Chomsky, N. (2002). *On nature and language*. Cambridge, England: Cambridge University Press.
- Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.
- Christiansen, M. H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, 23, 157–205.
- Culicover, P. (1999). *Syntactic nuts: Hard cases in syntax*. Oxford, UK: Oxford University Press.
- Culicover, P., & Jackendoff, R. (2005). *Simpler syntax*. Oxford, UK: Oxford University Press.
- Culicover, P., & Nowak, A. (2003). *Dynamical grammar*. Oxford, UK: Oxford University Press.
- Dennett, D. C. (1991). *Consciousness explained*. Boston, MA: Little, Brown.
- Dennett, D. C. (2013). *Intuition pumps and other tools for thinking*. New York, NY: Norton.
- Elman, J. (1990). Finding structure in time. *Cognitive Science*, 14, 179–211.
- Fodor, J. D. (1998). Parsing to learn. *Journal of Psycholinguistic Research*, 27, 339–374.
- Frank, S. L., Bod, R., & Christiansen, M. H. (2012). How hierarchical is language use? *Proceedings of the Royal Society B: Biological Sciences*, 279, 4522–4531.
- Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago, IL: University of Chicago Press.
- Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford, UK: Oxford University Press.
- Hale, K., & Keyser, S. J. (1993). On argument structure and the lexical expression of syntactic relations. In K. Hale & S. J. Keyser (Eds.), *The view from Building 20* (pp. 53–109). Cambridge, MA: MIT Press.



- Hale, K., & Keyser, S. J. (2002). *Prolegomenon to a theory of argument structure*. Cambridge, MA: MIT Press.
- Halle, M. (1973). Prolegomena to a theory of word-formation. *Linguistic Inquiry*, 4, 3–16.
- Halle, M., & Marantz, A. (1993). Distributed morphology. In K. Hale & S. J. Keyser (Eds.), *The view from Building 20*, 111–176. Cambridge, MA: MIT Press.
- Harley, H. (2014). On the identity of roots. *Theoretical Linguistics*, 40, 225–276.
- Hoffman, T., & Trousdale, G. (2013). *The Oxford handbook of construction grammar*. Oxford, UK: Oxford University Press.
- Jackendoff, R. (1975). Morphological and semantic regularities in the lexicon. *Language*, 51, 639–671.
- Jackendoff, R. (2002). *Foundations of language*. Oxford, UK: Oxford University Press.
- Jackendoff, R. (2007). A Parallel Architecture perspective on language processing. *Brain Research*, 1146, 2–22.
- Jackendoff, R. (2010). *Meaning and the lexicon*. Oxford, UK: Oxford University Press.
- Jackendoff, R. (2011). What is the human language faculty? Two views. *Language*, 87, 586–624.
- Jackendoff, R., & Audring, J. (2016). Morphological schemas: Theoretical and psycholinguistic issues. *The Mental Lexicon*, 11, 467–493.
- Jackendoff, R., & Audring, J. (forthcoming). Relational morphology in the parallel architecture. In J. Audring & F. Masini (Eds.), *Oxford handbook of morphological theory*. Oxford, UK: Oxford University Press.
- Joshi, A. (1987). An introduction to Tree-Adjoining Grammars. In A. Manaster-Ramer (Ed.), *Mathematics of language* (pp. 87–114). Amsterdam, North Holland: John Benjamins.
- Kay, P. (2013). The limits of (Construction) Grammar. In T. Hoffman & G. Trousdale (Eds.), *The Oxford handbook of construction grammar* (pp. 32–48). Oxford, UK: Oxford University Press.
- Lakoff, G. (1970). *Irregularity in syntax*. New York, NY: Holt, Rinehart & Winston.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211–240.
- Langacker, R. (1987). *Foundations of cognitive grammar*. Stanford, CA: Stanford University Press.
- Lees, R. B. (1960). *The grammar of English nominalizations*. The Hague, The Netherlands: Mouton.
- Lieber, R. (1992). *Deconstructing morphology: Word formation in syntactic theory*. Chicago, IL: University of Chicago Press.
- Marcus, G. (1998). Rethinking eliminative connectionism. *Cognitive Psychology*, 37, 243–282.
- Marcus, G. (2001). *The algebraic mind*. Cambridge, MA: MIT Press.



- Marr, D. (1982). *Vision*. San Francisco, CA: Freeman.
- Murphy, G. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.
- Pierrehumbert, J. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology VII* (pp. 101–139). Berlin, Germany: Mouton de Gruyter.
- Pinker, S. (1999). *Words and rules*. New York, NY: Basic Books.
- Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 26, 195–267.
- Pollard, C., & Sag, I. (1987). *Information-based syntax and semantics*. Stanford, CA: Center for the Study of Language and Information.
- Pollard, C., & Sag, I. (1994). *Head-driven phrase structure grammar*. Chicago, IL: University of Chicago Press.
- Prince, A., & Smolensky, P. (2004). *Optimality theory: Constraint interaction in generative grammar*. Oxford, UK: Blackwell.
- Pullum, G. (2013). The central question in comparative syntactic metatheory. *Mind and Language*, 28, 492–521.
- Rumelhart, D., & McClelland, J. (1986). On learning the past tense of English verbs. In J. McClelland, D. Rumelhart, & the PDP Research Group, *Parallel distributed processing* (Vol. 2, pp. 216–271). Cambridge, MA: MIT Press.
- Sadock, J. (1991). *Autolexical syntax*. Chicago, IL: University of Chicago Press.
- Searle, J. (1995). *The construction of social reality*. New York, NY: Free Press.
- Seidenberg, M., & McDonald, M. (1999). A probabilistic constraints approach to language acquisition and processing. *Cognitive Science*, 23, 569–588.
- Shieber, S. (1986). *An introduction to unification-based approaches to grammar*. Stanford, CA: Center for the Study of Language and Information.
- Siddiqi, D. (forthcoming). Distributed morphology. In J. Audring & F. Masini (Eds.), *Oxford handbook of morphological theory*. Oxford, UK: Oxford University Press.
- Spencer, A. (2013). *Lexical relatedness*. Oxford, UK: Oxford University Press.
- Stump, G. (1990). *Inflectional morphology: A theory of paradigm structure*. Cambridge, England: Cambridge University Press.
- Taylor, J. (2012). *The mental corpus*. Oxford, UK: Oxford University Press.
- Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.
- Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model, *Cognition*, 92, 231–270.
- Wasow, T. (1977). Transformations and the lexicon. In P. Culicover, T. Wasow, & A. Akmajian (Eds.), *Formal syntax* (pp. 327–360). New York, NY: Academic Press.

## 4.2 REFLECTIONS ON RAY JACKENDOFF

Daniel C. Dennett

There is a prevailing presumption that the *technical terms* of a discipline are more rigorously defined, more regimented, than the loosely anchored language of the “layman,” but this presumption is often false. Sometimes, a technical term gains popularity in a field precisely because it is *not yet* pinned down, permitting some waffling, some strategically noncommittal hand-waving to fill in, temporarily, for the precision that will be achievable eventually, when we come to understand the phenomena better. Think of *gene*, for instance, or *stimulus* and *response*. It is not that no effort was made to define these terms precisely when they were introduced but that a variety of different definitions have been permitted to compete, sometimes for decades. Ever since Socrates, philosophers have typically deplored this promiscuity, demanding that participants define their terms and refusing to engage in the substantive questions until consensus on definitions has been reached. Such policing of boundaries has on occasion led to valuable clarifications, clearing the ground for formal theories and models, for instance, but this semantic tidiness as often curdles into obstructionism, or the Heartbreak of Premature Formalization. Perhaps no technical term has been bandied about more casually than *mental representation*, with dozens of different senses proliferating since the demise of behaviorism opened the floodgates, and “cognitive science” rushed in. The result has been a period of tremendous growth of insight into how the brain must work its magic. The curmudgeons of (noncognitive) neuroscience who sneer at the fantasies indulged in by their cognitive colleagues should remind themselves of how the spectacular advances of molecular genetics would have been unimaginable without the ladder of the noncommittal concept of a gene that many have now discarded.

Ray Jackendoff's chapter provides a masterful guide to the current understandings of (*mental*) *representation* in linguistics, and makes the case for a way forward—his way—that promises to achieve some of the unification with biology that should be the goal of all cognitive science. For those of us who have neither the training nor the stamina to penetrate the thickets of contemporary controversy in linguistics, Jackendoff's periodic progress reports (1993, 2002, 2011, 2012), offer a refreshing perspective, written in user-friendly terms that permit the reader to identify the author's (authoritative) bias and assess it. Jackendoff has earned the right to express strong opinions, and he exercises it judiciously.

As he notes, some philosophers use "whole brain states" identified as, say, beliefs, as their mental representations (which is just treating a person as an intentional system and attributing beliefs to her), but once we begin entertaining hypotheses about how the subpersonal machinery of belief generation and language processing actually work, we trade in such bird's-eye view formulae for specific "data structures that pertain to a particular domain of brain activity." Note that this link is not (necessarily) to a specified region or module, let alone a specified neuron or cluster of neurons, but to a domain of activity, such as phonology, or an interface between two domains.

"For instance, one cannot just 'send' phonological representations to semantics. Phonological representations per se are completely unintelligible to semantics; and retinal images per se mean nothing to 'higher-level' visual representations such as those involved in face recognition. Hence, even if one notates the interface with an arrow, it has to signify a complex correspondence or conversion between one sort of data structure and another, and this conversion is an essential component of mental computation" (p. 98).

These posited representations are not *for* the people in which they are active, in the sense of being appreciated, understood, noticed by them; they are *for* the sub-personal subsystems that need them to accomplish their tasks. Jackendoff says they are "cognitively real" but "not physically real in any useful sense" (p. 97). In line with Clark's essay and my response, we might do better to set aside traditional questions about what is "real" and what is "illusory" and recognize that these are affordances (see also Kukla's essay and my response) of a rather special sort: real *patterns* that make it possible for the subpersonal machinery to do its work.

When we turn from the metaphysical status to the engineering powers of these representations, Jackendoff shows how his Declarative model avoids some of the artifactual constraints of the Procedural alternative (such as "directionality" and strict sequence) while providing the sort of productive structural discipline the No Rules model lacks. I'm not fond of these terms, but as labels they are slightly more informative than Type A, Type B, and Type C models. To me, the

main message is that he can move away from the apparently austere and timeless (unchanging) perspective of the generative linguistics of a few decades back and toward a more dynamic, evolving, brainish set of regularities, uniting (I would urge—see Clark, chapter 7.1, and my response, chapter 7.2) with the Bayesian models. Then the constructing of sentences on the fly begins to look more like the biochemical construction of macromolecules—proteins, enzymes, and the like, following “rules” of accretion.

This paves the way for Jackendoff to venture a new evolutionary hypothesis:

The process that upgrades a variable from closed to open is not omniscient. It has no way of knowing in advance whether a pattern is going to be fully productive and or only partially productive . . . If the hypothesis is validated, the relevant variables in the schema are upgraded to open. But if it fails, that does not invalidate the original assessment that a pattern exists. Hence alongside the fully productive schemas, one should expect to find lots of partially productive schemas that have failed to be promoted. Since such schemas are capturing partial regularities, there is no reason to discard them. We should expect the lexicon to be littered with failed (but still not useless) hypotheses. (p. 121)

Jackendoff is thinking of individual learning, but I want to consider it as nothing less than a gradualistic evolution (memetic, not genetic) of linguistic complexity, a “missing link” largely unimagined by linguists, who have been strongly discouraged from considering such Darwinian possibilities.<sup>1</sup> Can these predicted fossil traces be found? A delicious prospect, which might finally subvert the complacency of those who see no role for Darwinian thinking in cognitive science (see Baker, chapter 11.1, and my commentary, chapter 11.2).

An early passage in the essay provoked in me a conviction—well, a supposition—that Jackendoff might not share:

In the domain of meaning, the relevant mental representations differentiate concepts into objects, events, properties, places, and so on; the objects are differentiated into animate and inanimate and characterized in terms of shape and function; and so on down to the concept of a cat somewhere in the taxonomy, itself further differentiated into kinds of cats and cats one has known. (p. 97)

---

1. Jackendoff is a rare exception. See in Jackendoff (2002) his discussion of linguistic “fossils”—remnants of protolanguage features still to be found scattered in modern languages (pp. 236–256). See also Jackendoff and Wittenberg (2014), on the evolution of sign language.

It struck me that this involuntary categorization—which is, I think, a close kin of the automatic digitization one sees in phonology—is perhaps the birthplace of essentialism. That is, seeing things as having essences is perennially attractive, even after one has been given the opportunity to know better, and it might be the natural byproduct of a genetically installed proclivity to make fish-or-cut-bait mechanisms that impose a categorization scheme on our perceptual systems in the interests of practicality. Perception as triage, you might say.

## Works Cited

- Jackendoff, R. (1993). *Patterns in the mind: Language and human nature*. New York, NY: Basic Books.
- Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford, UK: Oxford University Press.
- Jackendoff, R. (2011). What is the human language faculty? Two views.” *Language*, 87, 586–624.
- Jackendoff, R. (2012). *A user's guide to thought and meaning*. New York, NY: Oxford University Press.
- Jackendoff, R., & Wittenberg, E. (2014). What you can say without syntax: A hierarchy of grammatical complexity. In F. Newmeyer & L. Preston (Eds.), *Measuring grammatical complexity* (pp. 65–82). New York, NY: Oxford University Press.



## CONSCIOUS EXPERIENCE



# 5.1 SEEMING TO SEEM

David Rosenthal

## 1. First and Third Person

An apparent perplexity about the mental is that we know about mental states—thoughts, perceptions, sensations, and so forth—in two ways, and these two ways of knowing about mental states seem to have little, if anything, to do with one another. We know about our own mental states by first-person access, at least when those states occur consciously. Their being conscious, by itself and with no input from any other source, seems to tell us not only which states occur in one, but also what the mental nature of those states is.

The seemingly unmediated conscious access we have to our own mental states has led many to see the mental as resisting any sort of objective, scientific treatment, and perhaps on that account as being nonphysical. And because conscious first-person access to mental states seems subjectively to be the last word about their nature and occurrence, many have been tempted to see such access as also having a privileged epistemic status, as being incorrigible or even infallible.<sup>1</sup>

But we plainly also often know about the mental states of others. We can often tell when somebody is in pain, sees something red, wants a beer, or thinks it's going to rain. Such third-person access is plainly not incorrigible, much less infallible; but it's often very good. We sometimes see that somebody else is in a particular mental state even before that person is aware of being in that state, if indeed the person ever does.

Indeed, our knowledge about others' mental states is sometimes good enough to override another person's subjective sense of what

---

1. The two are not the same; first-person access might not be subject to being corrected even if it is not always correct.



mental state that person is in. You may sometimes know better than I what I want or even what I'm thinking or paying attention to. And we sometimes know better than another what emotional state that person is in. There is even reason to think one can be mistaken about whether one is in pain (e.g., Koyama, McHaffie, Laurienti, & Coghill, 2005). First-person access to one's own mental states is not only not infallible; it is sometimes actually corrigible by appeal to others' observations.

Because first-person access to one's own conscious states and third-person access to others' mental states operate in such different ways, it can be difficult to do justice to both and, in particular, to know how the two ways of knowing about an individual's mental states fit with each another. How can the subjective grasp of one's own mental states square with the observation- and inference-based knowledge we have of the mental states of others?

One response might be to deny that first- and third-person access are about the same things. But this is unworkable. When you take me to be in some mental state, the state you take me to be in is typically of the same type as states that you are sometimes in yourself and have first-person access to. One indication that the states we ascribe in first- and third-person cases are of the same type is that if I take you to be in some mental state relying on third-person access and you, relying on your first-person access, deny being in that state, we are plainly contradicting one another. You are denying being in the very state that I assert you to be in. First- and third-person access are about states of the same type. We do not ascribe to others diminished or surrogate versions of the mental states we take ourselves to be in; the mental states we take others to be in and ascribe to them are the genuine article, states of just the sort that sometimes consciously occur in us.<sup>2</sup>

A condition on any acceptable account of mind, accordingly, is to square these two types of access we have to mental states, to explain how the states we have first-person access to are states of the very same type as those we also often know others to be in by third-person observation and inference. An account of the nature of mental states must make clear how those states are subject to both kinds of access.<sup>3</sup>

---

2. Another strategy, recently popular, would be to say that we know in a first-person way about mental states, at least qualitative mental states, by applying special phenomenal concepts, whose modes of presentation are the mental qualities actually present to one (e.g., Alter & Walter, 2007). But that mode of presentation would allow application of such concepts only to oneself, and no way has been given by proponents of this strategy to calibrate that application with concepts that apply across individuals. The problem about first and third person hasn't been solved, but just relocated from kinds of access to corresponding kinds of concept.

3. Though this thought has been downplayed somewhat in the recent literature, it is hardly new. See, for example, Strawson (1959, p. 104): "There would be no question of ascribing one's

The natural strategy is to begin with an account of mental states geared in the first instance to one kind of access and then show how that account can also accommodate access of the other kind. But that is easier said than done. First-person access presents itself subjectively as being wholly unmediated. Third-person access, by contrast, is plainly mediated by observation and inference, though the mediating observations and inferences are not always consciously explicit. So if we take this apparent lack of mediation at face value, any disparity between first- and third-person access should always be settled in favor of the first person. Mediation always allows for the introduction of error; though lack of mediation doesn't guarantee accuracy, it arguably wins in any contest with mediated access.

So if we take the apparent immediacy of first-person access seriously, we may see whatever inferences one might make about others' mental states as mere informal guesses as compared to first-person pronouncements. But we know that first-person access is on occasion less accurate than observation and inference by others. Giving pride of place to the first person unavoidably shortchanges the robust third-person access we do actually have to others' mental states.

It's also not clear what kind of state could be subject to wholly unmediated access. Some have been tempted by the thought that our access to abstract objects, such as numbers, is wholly unmediated; but how could we know about concrete states of ourselves in way that is absolutely unmediated? And how could states accessible in an absolutely unmediated way also be accessed in a third-person way? Third-person access would of course rely on causal ties the states have with observable occurrences, such as behavior. But first-person access seems to reveal no causal connections the states have with anything observable; indeed, such access, by itself, seems not even to represent mental states in respect of causal ties with one another.<sup>4</sup>

So taking first-person access as primary undermines not only our understanding of how mental states could be accessible in a reasonably reliable way from the third person, but also the importance to the nature of those states of their causal ties with behavior, sensory stimulation, other mental states. It is doubtless this

---

own states of consciousness, or experiences, to anything, unless one also ascribed, or were ready and able to ascribe, states of consciousness, or experiences, to other individual entities of the same logical type as that thing to which one ascribes one's own states of consciousness. The condition of reckoning oneself as a subject of such predicates is that one should also reckon others as subjects of such predicates."

4. We do, of course, assume causal ties that mental states accessible from the first person have with behavior and with other mental states, but we infer such causal ties by applying folk-psychological generalizations, not because first-person access reveals them. I'm grateful to Pete Mandik for raising this concern, and for a helpful reading of an earlier draft.

that leads many who favor starting from the first person to downplay or sometimes even ignore altogether the need for an account to accommodate both types of access, and to deny that causal ties are essential to the nature of mental states.

It is partly for these reasons that Daniel Dennett's work on the mind has been so important and influential and has had such a deeply salutary effect. Dennett's (1978, 1987, 1991b) appeal to the intentional stance holds that we ascribe to others the intentional states that make sense on an assumption of those individuals' rationality. It thereby paves the way for a full and rich account of mind that starts not from first-person access to mental states, but from the way we know about those states from the third person. And it helps undermine the now popular claim that a satisfactory account of mind must rest primarily, or possibly even exclusively, on first-person access. We can understand the mind by appeal to ascriptions of mental state based on behavior and context.<sup>5</sup> Dennett's endorsement of the primacy of a third-person approach echoes in this way arguments of W. V. Quine (1960, esp. section 45; 1985) and Wilfrid Sellars (1956, 1968).

It is important to note that the assumption of rationality that sustains the intentional stance does not require that one take all the intentional states one sees an individual as being in as rational, even given that individual's purposes and proclivities. The intentional stance applies to individuals some of whose intentional states depart to some extent from full rationality. But one cannot understand an individual as having thoughts and desires at all unless one takes those states to be rational on balance. The rationality assumption is in that way of a piece with a principle of charity that governs how we understand one another's speech. On that principle, each of us takes the speech acts of others to be true as much as possible, to embody valid inferences as much as possible, and to use words in standard ways as much as possible.<sup>6</sup>

Starting from first-person access makes it difficult to do justice to the way we have access to others' mental states from the third person and the way that access fits with our first-person access. This is in part because first-person access seems subjectively to be the last word about the nature of mental states and in part because such access seems to represent the nature of those states independent

---

5. As Dennett stresses, use of the intentional stance is appropriate only when one can't explain the behavior of something by the less elaborate assumptions that figure in the design or physical stances.

6. Maximizing of truth, validity, and standard use of words must be by the lights of the individual doing the interpreting; there is no other way. And since it may not always be possible to maximize all three, one will sometimes have somehow to strike a balance. Because there are often alternative ways to balance the maximizing of the three, there will inevitably be conflicts of interpretation that cannot be settled in any independent, objective way. See Rosenthal (1989, section 4; 2009, section 6).

of the causal ties those states have with behavior, stimuli, and other mental states. So perhaps starting from the third person holds greater promise for an account that does justice to both kinds of access.

But those who favor the primacy of the first person will push back, urging that we will never be able to do justice to first-person access if we start instead with what it is about mental states that enables third-person access. So they will argue that we face a kind of standoff; according primacy to either type of access makes doing justice to the other and how the two fit together difficult if not impossible. Since we cannot do full justice to both, we must choose. And given that choice, they will conclude, first-person access plainly wins, since that is what is unique to and distinctive of the mental. We must accordingly give pride of place to the first person.

But that is not the choice we face. The apparent failure of first-person access to represent mental states in respect of at least many of their causal connections does likely doom any attempt to start with the first person and arrive at a satisfactory account of third-person access or, indeed, a full account of the nature of those states themselves. But the advocate of the primacy of the first person is arguably wrong that the opposite is also the case. We can build on a third-person account of the mental to arrive at an accurate and full explanation of the first-person aspect of mental states and of our conscious access to those states. We can usefully see Dennett's *Consciousness Explained* (1991a) and his extensive subsequent work on consciousness (e.g., 1996, 2015, and esp. 2005, chapter 2) in this light, as building a satisfactory account of first-person access and the first-person aspect of mental phenomena that rests on third-person considerations that fit with those that animate the intentional stance approach in (1987).

Evaluating this is of course a delicate matter. Can an account that fits with the third person deliver everything an advocate of a first-person approach would require? In particular, can an account that fits with an intentional-stance deliver those goods? Moreover, is everything that those who favor a first-person approach demand warranted? Might some demands that those who favor the first person make simply be results of privileging the first person, and not independently motivated or warranted?

As we have seen, similar questions arise in evaluating whether a first-person approach can deliver a full and satisfactory account of the mind. Is first-person access subjectively privileged in the way it seems? Are the causal ties between mental states and observable occurrences and among the states really essential to the nature of those states?

Champions of the first person will urge that we accept the subjective privilege as true, and deny that causal ties are essential for a full account of mental states. It would then be reasonable to see third-person access as not that important to

accommodate. The claim that causal ties are merely contingent to the mental states receives vivid expression with the intriguing idea that the qualitative states that occur in perceiving might be undetectably invertible from one individual to another. Perhaps, on this line, the mental quality that occurs when you see something red, for example, is the same type of mental quality that occurs when I see something green.

Such undetectable inversion would be possible—indeed conceivable—only if the causal ties those states have with observable stimuli were inessential to those states. And that would impair the kind of access we can have to others' mental states. One could only know that when somebody else sees something red the person is in the kind of state distinctive for that person when seeing red; one would not also know what kind of state it is, described in mental terms. Only first-person access could then reveal the nature of qualitative mental states; third-person access says nothing informative about the distinctively mental nature of others' states. And if causal ties are inessential to intentional states as well, similar considerations would apply there; one could know the nature of those only from the first person.

By contrast, a view that accords primacy to third-person access rejects the possibility—indeed, even the conceivability—of such undetectable inversion. And our folk picture of mental states concurs. Causal connections with stimuli are not merely incidental to mental states, but an aspect of their mental nature; here the primacy of third-person access coincides with common sense. Our folk view holds that we ascribe to others states of the very same type we are aware of as occurring in ourselves; so that view prevents us from even conceiving of the mental states others are in as undetectably inverted from our own.

Some will insist that even if undetectable inversion isn't possible, it is at least conceivable (e.g., Shoemaker, 2003, p. 336). But undetectably invertible states could not be mental states as our folk conception dictates. Compare conceiving of something observably exactly like water, though it isn't  $H_2O$  (Rosenthal, 2010, sections 1, 3). We are in that case not conceiving of water, even if the imagined substance is superficially just like water.<sup>7</sup> By the same token, conceiving of undetectably invertible states is not conceiving of mental states. The pretense that we can imagine or conceive of undetectable inversion of mental states stems from the playful conceit that first-person access is all that matters—indeed, all that could conceivably matter—in determining the nature and occurrence of mental states. But we know on the slightest reflection that this isn't and couldn't be so.

---

7. On cases like this see Kripke (1980, pp. 130–140, esp. 130–134; cf. also p. 104); Kripke relies primarily not on the example of water, but on heat's being mean molecular kinetic energy.

All this has important consequences for evaluating the dueling strategies of starting with the first or third person in constructing an account of the mental on which both types of access fit together. If the champion of the first person argues that the conceivability of undetectable inversion is a pretheoretic datum that an account of mind must preserve, the natural reply is that it is not a datum at all, but at best an artifact of taking the first person as our starting point. Similarly with the alleged incorrigibility or infallibility of first-person access; these are not pretheoretic, commonsense data, but just the result from taking the first person by itself to reveal the nature of mental states. In evaluating an account that begins from the third person, we must exclude demands that rely only on taking first-person access to be our only genuine access to the mental.

The champion of the first person often appeals to undetectable invertibility, incorrigibility, and infallibility as pretheoretic intuitions about mental phenomena that any account must accommodate. But such alleged intuitions are typically just appealing ways to package substantive theoretical claims, passing them off as pretheoretic data that impose constraints on theorizing. Adapting Dennett's useful notion of an intuition pump,<sup>8</sup> we can think of these so-called intuitions as *theory pumps*: devices to get us to adopt a tacit but controversial theoretical approach.

No such theory pumps would be needed, and so none would be invoked, were there independent, non-question-begging support for the theoretical approach in question. Here the alleged intuitions simply package the theoretically contentious and otherwise unsupported view that first-person access exhaustively reveals the nature of mental states.

But even if we took intuitions at face value, not as theory pumps but as genuinely pretheoretic convictions, there would still be two ways to accommodate them. We could, as the advocate of first-person primacy urges, take the convictions as imposing constraints on any satisfactory account. That would be to save the intuitive appearances by taking them to be accurate. But we need not accommodate intuitions in that way; we could instead simply seek to explain why we have the convictions in question and why they seem so compelling.

The second approach does not require assuming that the convictions are true; there are many compelling convictions that turn out not to be true. And even if they were true, we would need to explain why we have them. The traditional

---

8. Dennett originally coined the term "intuition pump" to describe "thought experiments . . . [that involve] inviting the audience to imagine some specially contrived or stipulated state of affairs, and then—without properly checking to see if this feat of imagination has actually been accomplished—inviting the audience to 'notice' various consequences in the fantasy" (1991a, p. 282). Also: "an intuition pump—a story that cajoles you into declaring your gut intuition without giving you a good reason for it" (1991a, p. 397). See also Dennett (2013).

explanation smuggles in the controversial assumption that the mind is transparent to itself; the reason we have the intuitions is because such transparency leads automatically from the nature of mind to true intuitions about the mind. But assuming such transparency is patently question begging in the context of evaluating the intuitions. The very appeal to intuitions about the mind is simply a way to package the question-begging theoretical assumption that we must rely only on first-person access to learn about the nature of mind.

In what follows, I examine how well Dennett's treatment of consciousness based largely on third-person considerations does justice to first-person access and saves those first-person appearances that we have independent reason to respect. First-person access and the appearances it generates are a matter of mental states' being conscious; so evaluating how well an account saves legitimate first-person appearances hinges on how that account deals with consciousness. In section 2, I briefly sketch Dennett's account of consciousness. In section 3 I raise some difficulties for that account. I conclude in section 4 by showing that these difficulties point toward an alternative theory of consciousness, strongly similar in spirit to Dennett's but with some important differences.

## 2. Dennett and Consciousness

Key to Dennett's views about consciousness is a type of problem case he raises about timing (1991a, chapters 5, 6; Dennett & Kinsbourne, 1992). It will sometimes happen that one has a conscious visual experience that is inaccurate due to the interference of a memory one has of a similar or associated object or scene. In Dennett's vivid example, you see a long-haired woman without eyeglasses running by, but the memory of having seen a different short-haired woman with eyeglasses intrudes, and your conscious experience is of seeing a short-haired woman with eyeglasses (1991a, p. 117–118). Perhaps the memory is of a friend and, though you don't know the passing woman, you consciously experience seeing her as your friend. We have all doubtless been subject to this kind of thing; memory does sometimes skew perception in this way.

The memory of your friend enters the causal stream relevant to your conscious experience and distort the visual information sometime between retinal stimulation and stable conscious experience. But exactly when? One possibility is that it affects things before the perceptual input becomes conscious. But there is another possibility. Perhaps you first have an accurate conscious experience of the passing woman as having long hair and no eyeglasses, but the memory immediately intrudes, and your initial conscious experience is immediately replaced by an inaccurate conscious experience of a short-haired woman with eyeglasses. The initial conscious experience, moreover, is so brief that you have no memory of it;



subjectively it's just as though you'd never had that initial conscious experience. The two types of occurrence, resulting from interference by a memory at different points, would be subjectively indistinguishable.

Folk psychology seems to offer nothing that favors one possibility over the other. Dennett colorfully labels the first possibility *Stalinesque*, since like Stalin's show trials consciousness presents us with false information; he calls the second *Orwellian*, because in that case consciousness would in effect rewrite history. Since neither subjective impressions nor folk psychology can help, Dennett sees no way to determine in any particular case when it is that memory contaminates the visual information, before the perceptual input comes to be conscious (*Stalinesque*) or after (*Orwellian*).

Dennett also surveys a variety of striking experimental results in which timing seems to pose a similar problem about conscious experience. In the *phi* phenomenon, two round patches of light appear briefly one after the other. If the spatial distance between them and the timing of their successive blinking are just right, observers don't experience two disks lighting up in succession, but rather a disk moving back and forth between the two locations. And if the stimuli differ in color, the moving disk appears to change color in midstream (Dennett 1991a, p. 114; Kolers et al., 1976; Nelson Goodman suggested testing for color change).

But the first blinking of the first light would not by itself result in the illusion of a moving disk. So what happens after that first blink stimulates the retina, sending visual information onto visual cortex, prior to the blinking of the second light? Does the visual input of the first light by itself simply never reach consciousness (*Stalinesque*)? Or does it become conscious only to be replaced so fast by a conscious experience of a moving disk that no memory remains of the conscious experience of the blinking of the first light by itself (*Orwellian*)? Again, neither subjective impressions nor folk psychology helps settle the question. Other experimental results pose similar quandaries.

Dennett argues that these phenomena have important consequences for understanding consciousness. If there is no nonarbitrary way to settle whether the initial perceptual input in these cases makes it to consciousness, it's to that degree indeterminate as to whether particular perceptual inputs occur consciously. And that, he argues, casts doubt about how determinate consciousness is in general. He concludes that "there is no reality of consciousness independent of the effects of various vehicles of content on subsequent action (and hence, of course, on memory)" (1991a, p. 132).

In the running-woman case, the visual input and the memory of one's friend compete with one another as to which will make it to and remain in consciousness. Similarly with the first blink and the illusion of motion generated by



alternating blinks. So Dennett urges that we mustn't think that there are a number of well-defined psychological states, each of which may in the ordinary course of events make it to consciousness. Rather, there are a number of psychological drafts, often conflicting, and each competing with many others to make it to consciousness. This is Dennett's Multiple Drafts model of consciousness.

This conclusion is not all that surprising. It is natural to think that the mind involves many competing causal strands, and these play out in determining what makes it to consciousness. But Dennett also draws a more contentious conclusion from the puzzles about timing. If there is no way to tell whether a particular case occurred in a Stalinesque or Orwellian way, he concludes, then "there are no fixed facts about the stream of consciousness independent of particular probes" (1991a, p. 138; cf. p. 275). Such probes may include somebody's asking what one perceives or the demands of action or the like making it relevant what one consciously perceives. The facts of consciousness are fixed by the effects consciousness has on other things. It is this statement of his view that Dennett labels "first-person operationalism" (henceforth "FPO"; 1991a, p. 132).<sup>9</sup>

Opposition to operationalism in general, Dennett notes, stems from the idea that there are facts not caught by whatever operationalist test one might propose. But operationalism about consciousness, he argues, is special; who would maintain that there are facts of consciousness not subject to first-person access? We must, he insists, deny the possibility in principle of consciousness of a stimulus in the absence of the subject's belief in that consciousness" (1991a, p. 132). In the cases described above, if the difference between a Stalinesque and an Orwellian occurrence would make no subjective difference to the individual in question and they don't in any other way differ in outcome, there is simply no distinction to be drawn. The appearance of a distinction here is illusory. Operationalism in respect of an individual's beliefs catches all the facts there are about consciousness.

It might not seem to matter all that much whether we can determine which of the Stalinesque and Orwellian mechanisms is operative in generating a particular case of conscious misperception, or even whether we can draw a tenable distinction between them. But FPO also rules illusory another distinction that figures more deeply in theorizing about consciousness. Dennett takes "writing it down"

---

9. The idea that competition among many nonmental factors results in some folk-psychological conscious states may encourage another claim of Dennett's, that the search for a neural correlate of consciousness "is probably a wild goose chase" because "the processes that elevate contents to consciousness are like the processes that elevate [evolutionary] lineage divergences into speciation events" (2009, p. 234). But the analogy with evolution is questionable; the competing factors may well result in states we can taxonomize along standard folk-psychological lines, allowing for a subsequent relatively specific process that results in those states' becoming conscious.

in memory [to be] criterial for consciousness; that is what it is for the 'given' to be 'taken'—to be taken one way rather than another" (1991a, p. 132). Perceptual input is conscious if, but only if, it's taken in some way; that's why there is no "consciousness of a stimulus in the absence of the subject's belief in that consciousness." The distinction between how consciousness is and how it seems is illusory. That's the operationalism in first-person operationalism.

So on FPO, one cannot be wrong about the facts of consciousness in one's own mental life; how consciousness seems to one is how it actually is. This may initially seem right, perhaps even obvious. Consciousness is itself just a matter of seeming; there is nothing to it except seeming. And this may seem to amount to endorsing the privilege that traditional theorists have accorded first-person access. If how it seems is how it is where consciousness is concerned, then one cannot be wrong about one's own conscious states. Dennett lampoons the view he calls the Cartesian Theater, on which consciousness consists the viewing by a metaphorical viewer of a stream of consciousness. But in adopting FPO, it appears that he may in effect be accepting a version of the traditional Cartesian view that first-person access is infallible.<sup>10</sup>

The apparent distinction between Stalinesque and Orwellian mechanisms is an important opening wedge for Dennett. The reason there is, on his view, no fact of the matter of any sort about which mechanism is operative is that there is no subjective fact of the matter; which mechanism is operative would, by hypothesis, make no difference to how things seem subjectively to the individual. That's simply the way the alternative processes are described. The illusory character of a distinction between Stalinesque and Orwellian is of a piece with the view that how it seems for consciousness exhausts how it actually is. There being no way to determine which mechanism is operative reveals, Dennett urges, that there is nothing to consciousness beyond what the individual in question thinks there is.

And because how it seems for consciousness is how it is, there is no room for what Dennett stigmatizes as "the bizarre category of the objectively subjective—the way things actually, objectively seem to you even if they don't seem to seem that way to you!" (1991a, p. 132). Its seeming to one that it seems a particular way cannot amount to anything beyond its simply seeming that way to one. If one consciously sees a red square, it seems to one that there's a red square over there. What could it be, Dennett is asking, for it to seem to one that it seems to one that there's a red square over there? What could it be, that is, apart from its simply seeming to one that there's a red square over there?

---

10. As Fred Dretske (2000, p. 138) wryly notes: "First-person operationalism sounds like warmed-over Cartesianism to me."

If its seeming that it seems that there's a red square could differ from its simply seeming that there's a red square, we could distinguish the reality of seeming from how that seeming appears, how subjective seeming objectively is from how that seeming appears to one. So Dennett's denial that there are such levels of seeming, and hence a difference between how a case of seeming really is from how it appears, may strike one as straightforward common sense. There is nothing to how seeming appears, one may want to insist, beyond the seeming itself, no distinction here between appearance and reality. As Thomas Nagel puts it, when it comes to conscious experience "the idea of moving from appearance to reality seems to make no sense" (1974, p. 444).

But Dennett's denial of levels of seeming and his insistence that there are no facts about consciousness apart from the states one believes one is in do not rest on a first-person approach to the mind such as that championed by Nagel. The reason we cannot distinguish seeming to seem from mere seeming, according to Dennett, is not that we have special access to our own mental states to provide us with infallible self-knowledge. Rather, it is that there simply is nothing to our conscious experiences beyond what we believe there is. That is why we cannot be wrong about them, and why the idea of seeming to seem can, according to Dennett, get no purchase. Dennett has in effect repackaged the traditional Cartesian idea of infallible access in an operationist approach to the mind, thereby removing its epistemic bite.

To the extent to which we can see Dennett's view as involving access to one's own mind at all, it is a feature of his operationalism about the mind. It does not rest on something special about how the mind works or on how we have access to our inner mental workings. It is simply that our beliefs about our conscious experiences are the last, and indeed the only, word about those conscious states. We aren't wrong about our conscious experiences because when it comes to conscious experiences there is nothing beyond our beliefs to be wrong about.

We will see reason to question whether such operationalized first-person privilege is tenable. But for now, it's important to stress that the operationalist feature of Dennett's account allows him to construct a treatment of consciousness and first-person access that rests firmly on third-person considerations. Consciousness occurs, on Dennett's view, only if being conscious of something has an effect on memory, speech, or action, and the effect on memory is describable in speech. The multiple drafts of our psychological functioning get fixed as conscious states only once they can have discernible effects that are subject to interpretation.

One can adopt the intentional stance not only toward others, but also towards oneself. So it may seem inviting to see Dennett's view as holding that consciousness is a matter of self-interpretation, a matter of one's tacitly adopting the intentional stance towards oneself. This would fit comfortably with

Dennett's denial of a metaphorical internal viewer of a stream of consciousness. As already noted, consciousness on Dennett's view is "what it is for the 'given' to be 'taken'—to be taken one way rather than another." And it is natural to see this taking as a kind of self-interpretation. It is self-interpretation on the intentional-stance assumption that one is on balance rational. And if Dennett's theory of consciousness rests in this way on the application to oneself of the intentional stance, it is an account of consciousness built solidly on a third-person approach to the mind.

As Dennett stresses, the intentional stance is thoroughly realist about the occurrence of mental states. There are patterns of behavior that one can discern only by adopting the intentional stance toward an individual; these patterns are real, and they would be missed without the intentional stance. The real patterns sustain the realist character of the intentional-stance account of mind (1987, chapter 2, postscript; 1991b).

So the intentional stance will be thoroughly realist when applied to oneself as well, in connection with consciousness. The states one interprets an individual to be in by using the intentional stance are states the individual is objectively in, though one will discern them only by adopting the intentional stance. The interpretivist character of the intentional stance does not undermine the independent objectivity of the patterns one discerns with that stance; Dennett's interpretivism readily accommodates realism about the mind.

But FPO cannot be simply the application to oneself of the intentional stance. For one thing, applying the intentional stance involves the discerning of real patterns of behavior that allow one to ascribe beliefs and desires given a background assumption of overall rationality. And one plainly does not ascribe mental states to oneself on the basis of observing patterns of behavior. The states one does self-ascribe are indeed of the sort one can also ascribe to others by noting relevant patterns in their behavior; but observing such patterns does not figure in one's self-ascription.

Rather, there is on Dennett's view a competition among various internal states as to which will dominate one's behavior, including the speech behavior by means of which one announces what conscious mental states one is in. And he argues that there is nothing beyond the results of that competition that could settle what one's conscious states are. There is nothing, for example, that could settle whether the visual input of the running woman with long hair and no eyeglasses makes it to consciousness and is then revised or is revised before making it to consciousness. FPO is operationalist in simply dismissing that question, since there is nothing operational to determine an answer.

And FPO embodies a third-person approach to the mind since it allows nothing except what a person says as a way to determine what that person's

conscious states are. One will self-ascribe in response to a question's being put to one or some other probe, but "there are no fixed facts about the stream of consciousness independent of particular probes." One takes oneself to be in this or that state of consciousness, and that's the last and only word about it. So FPO is not simply a matter of applying the intentional stance to oneself, though it is in that spirit. Like the intentional stance we adopt toward others, FPO is interpretationist in that one's beliefs and reports of one's conscious states are the last word about what conscious states one is in. How one interprets oneself in respect of one's own conscious states settles the question of what one's conscious states are.

### 3. Seeming to Seem

How one takes one's conscious states to be is, according to FPO, the last word about the actual nature of those states. The operationalist character of FPO collapses any distinction one might seek to draw between what conscious states one is in and the way one takes one's conscious states to be, between one's conscious mental life and what one believes about one's conscious mental life. Dennett sees this as the right result. Consciousness is a matter of how things seem to us subjectively; so there can't be a tenable distinction between seeming and seeming to seem. If things seem to be a particular way, what could seeming to seem amount to other than simply seeming?

A satisfactory account of mind must do justice to both our first- and third-person takes about the mind, and it's likely that no account that takes first-person aspects of mind as primary can do the job. Starting from the first person is unlikely to succeed, in part because it deprives us of information needed to construct a satisfactory third-person account of mental functioning. Dennett's intentional-stance approach is grounded in the third person, and his FPO is designed to fit comfortably with that, taking one's own interpretation of one's conscious mental life as the first and last word about what one's conscious states are. In the absence of probes that result in one's taking oneself to be in particular conscious states, there is no fact of the matter about what conscious state one is in. So if FPO results in the collapse of any distinction between seeming and seeming to seem, we should, he would urge, accept that as the result of independently sound methodological considerations.

Those who favor an account of mind based on first-person access also tend to deny any distinction between seeming and seeming to seem. All seeming is intrinsically conscious on that approach. So consciousness cannot misrepresent our mental lives, and we are accordingly infallible about our mental goings on. And since we know of nothing else whose nature and occurrence we cannot be

mistaken about, there is a difficulty, perhaps intractable, in giving an objective account of the kind of subjectivity that conscious seeming involves.

Since Dennett's version of infallibility and subjectivity doesn't rest on anything special about the nature of mental states and consciousness, typical advocates of a first-person picture of mind would likely deny that Dennett saves the phenomena those advocates see as crucial. But such a complaint about Dennett's version of subjectivity and infallible access is arguably without foundation. The interpretativism of FPO does result in a type of subjectivity that stands apart from the rest of objective reality, and has first-person beliefs about one's conscious states be the last word about them. This arguably saves the substance of what advocates of the first person insist on, though it also dispels the intractable mystery inherent in an appeal to intrinsic consciousness.

But there is a different worry about Dennett's overall account, having to do with whether it enables first- and third-person access to fit comfortably together. Suppose that I believe that it's raining. Dennett's intentional stance should, at least on some occasions, enable you to tell that I have that belief. You adopt the intentional stance towards me, discern relevant patterns, and on that basis ascribe to me a belief that it's raining. That's all objective. You might make a mistake, but such a mistake, like any others about observable matters, can be corrected, and taking account of enough relevant patterns should do the trick.

But Dennett's FPO insists that "there are no fixed facts about the stream of consciousness independent of particular probes." How does that fit with your ability to tell that I believe that it's raining by discerning the relevant real patterns in my behavior? Does that mean that the intentional stance operates independent of "fixed facts about [an individual's] stream of consciousness"? Are the facts that can get fixed by particular probes independent of what the intentional stance enables one person to ascribe to another?

We can see how this difficulty arises by focusing on the claim of FPO that one's beliefs about one's own mental states are the last word about what they are. Suppose I believe that I believe it's sunny out and say that that's my belief. My pronouncements are the last word about what conscious beliefs and desires I have. But how can my belief and pronouncement be the last word if your adopting the intentional stance gives you third-person access to what beliefs and desires I have? Suppose the two deliver different verdicts. How does my having the last word square with your ascriptions grounded in third-person observation of my real patterns?

A natural answer to this quandary would be that the patterns you discern in my behavior enable you to ascribe beliefs and desires to me independent of whether those states are conscious, whereas my having the last word about what conscious states I'm in pertains only to my beliefs and desires only insofar as they

are conscious. Your observations can reveal the states I'm objectively in, whereas my beliefs about what conscious states I'm in is the last word only about how my subjective stream of consciousness appears to me. How my stream of consciousness appears to me is itself objective; it appears to me in one way and not another, and my beliefs about that are the last word about how it appears to me. But that objective matter is different from the objective matter your observations of real patterns reveal, namely, beliefs and desires I have independently of how my mental life appears to me.

But that way of squaring things is not available to Dennett; indeed, he emphatically rejects it. For me to be in particular intentional states is for things to seem a particular way to me. So if we distinguish the intentional states I'm in from the intentional states I am aware of myself as being in, that would be to distinguish how things seem to me from how they seem to me to seem. And that's the very distinction Dennett decries. Countenancing seeming to seem as distinct from merely seeming "creates the bizarre category of the objectively subjective—the way things actually, objectively seem to you even if they don't seem to seem that way to you" (1991a, p. 132).

The point is worth stressing. When you, using the intentional stance, ascribe beliefs and desires to me, you ascribe states that pertain to how things seem to me. And when the facts of consciousness are fixed by my beliefs about what I am conscious of, I too ascribe to myself states that have to do with how things seem to me. The problem was how your ascriptions can be objective if my self-ascriptions are the last word. How can there be "no fixed facts about the stream of consciousness independent of" probes that determine what states I take myself to be in if your intentional-stance ascriptions are objective?

The inviting solution is that my beliefs about what states I'm in fix only what conscious states I'm in, whereas your application of the intentional stance to me objectively reveals what states I'm in independently of their being conscious. Indeed, it's likely that this is the only way to deal with the potential conflict. But the difficulty with that for Dennett is that the states ascribed using the intentional stance are themselves states of seeming. So if states of seeming independent of consciousness are distinct from conscious states of seeming, there is after all a distinction between seeming and seeming to seem, which Dennett is at pains to deny.

We cannot get Dennett's intentional stance and FPO to fit well together unless we accept a distinction between seeming and seeming to seem. But there is reason independent of that difficulty to endorse that distinction. Perceiving often occurs consciously, but it also occurs without being conscious, as in subliminal perception. In typical experimental work in masked priming (Breitmeyer & Ögmen, 2006; Ögmen & Breitmeyer, 2006), a stimulus is briefly presented,



followed by another stimulus. If the target stimulus is presented on its own, subjects report seeing it, but when it is followed by a suitable mask they report seeing only the mask. Still, the target stimulus that subjects claim not to see influences subsequent psychological processing, and that is compelling evidence that the stimulus was seen, just not consciously.

Other cases involve stimuli presented to the blind field of a blindsight patient (Weiskrantz, 1997). Though the individual sincerely denies seeing the stimulus, elicited guesses are far above chance as to its shape and color and in some cases motion. This is so even with guesses about the emotional expression of faces presented to the blind field (de Gelder, Vroomen, Pourtois, & Weiskrantz, 1999). And there is nonconscious change detection; subjects report seeing no change, yet priming effects demonstrate that the change was seen, but not consciously (Fernandez-Duque & Thornton, 2000; Laloyaux, Destrebecqz, & Cleeremans, 2003).

It would be arbitrary and groundless simply to deny that seeing occurs in these cases. The visual input here has the same downstream psychological effects that are characteristic of conscious seeing, effects for example that vary with the color and shape of the stimulus and that affect desires. The only difference is that the input does not result in visual states the individual is aware of. But the visual input does register psychologically in the way distinctive of seeing; it simply doesn't register consciously. We would need some reason to deny that subliminal states constitute genuine perceiving. One could just dig in one's heels and deny that perceiving can occur without being conscious. But what reason could there be for that apart from traditional Cartesian claims that the mental is necessarily conscious, claims that are themselves without independent support?

Perceiving, moreover, involves things' seeming a particular way to one; if one is presented with a red, square stimulus, it will seem that something is red and square. Since perceiving can occur without being conscious, so can seeming. Consciously perceiving something involves one's being consciously aware of that thing, whereas perceiving the thing nonconsciously involves being aware of it, but not consciously aware of it.

So there is after all something to "the way things actually, objectively seem to you even if they don't seem to seem that way to you." If one subliminally sees a red, square stimulus, it will not seem to one that one sees that stimulus; one will sincerely deny that one does. Still, we can experimentally establish that the relevant area of one's visual field seems to one to have a red, square stimulus in it; one will guess if pressed to do so that there is a red square in that place, guesses that are far above chance. And one will behave psychologically in other ways as though one has just seen something red and square. So if one sees the stimulus



subliminally, there is a way in which it does seem to one that the stimulus is there even though one is unaware of its seeming that way to one.

But it will not seem to one that one sees a red, square stimulus, since one's seeing it isn't conscious. Things seem to one a particular way even though it does seem that they seem that way. There is after all room for a coherent "category of the objectively subjective—the way things actually, objectively seem to you even if they don't seem to seem that way to you."

Consciousness, according to Dennett, is "what it is for the 'given' to be 'taken'—to be taken one way rather than another" (1991a, p. 132). Subliminally seeing a red square is taking there to be a red square, though one is unaware of that taking. Being unaware of that taking, one takes oneself not to take there to be a red square. It seems to one that there is a red square; it's just that it doesn't seem to one that it seems that way. Since consciousness is a matter of such takings, there are evidently two levels of taking we must consider: how we take things to be and our awareness of our taking things to be that way. These two levels must be factored into any complete account of perceiving.

Dennett has argued against the occurrence of "qualia as traditionally conceived" (2015, p. 3; cf. 1991, chapter 12; and 1998, chapter 8). The traditional conception he has in mind is qualia as "unanalyzable simples"; as he notes, consciousness seems to present qualitative mental states in that way (2015, p. 8). But if qualitative states occur not only consciously, but also without being conscious, the idea that their qualitative properties are unanalyzable simples loses its force, as does the denial of such properties occur.

We seldom talk in everyday situations about perceptual or other mental states that aren't conscious. But that shouldn't tempt us to deny to the subliminal cases the status of genuine perception. We rarely talk about our own mental states when they aren't conscious because we are rarely aware of them; we would become aware of them only in some third-person way, say, by experimental results or by the application of the intentional stance to ourselves. We do sometimes note by observing others' behavior that they are in psychological states they are unaware of, though social niceties typically inhibit our commenting on that. But our tendency not to talk about perceptual states that aren't conscious doesn't show that there aren't any.

Beliefs also involve things' seeming to one to be a particular way. So even if one had qualms about genuinely perceptual subliminal states, it's widely accepted that one can believe various things without being aware that one does. A belief's not being conscious, moreover, does not prevent its being discernible from the intentional stance; so nonconscious beliefs are objective. Such nonconscious believing is another "way things actually, objectively seem to you even if they don't seem to seem that way to you." We cannot deny that seeming to seem occurs

that is distinct from mere seeming. And we accordingly cannot accept that what one takes one's own states of seeming to be is the last word about those states of seeming.

We use the term "conscious" in two ways, which it is important to distinguish. We speak of seeing, believing, desiring, and other mental states as being conscious; being conscious is a property we ascribe to mental states. But we also speak of seeing, believing, and many other mental states as states of being conscious *of* things. Here being conscious of something is the same as being aware of it; it's not that the state is conscious, but that being in the state is a way of being conscious or aware of something.

Subliminally seeing a red, square stimulus is being subliminally conscious or aware of that stimulus; one is aware of the stimulus, just not consciously aware of it. It is a way of being aware of the red square, even though the state of seeing is not on that account a conscious state. Being conscious of something can dissociate from the relevant psychological state's being a conscious state.

If there were no seeming to seem distinct from merely seeming, we couldn't distinguish conscious from subliminal perceiving, nor conscious from non-conscious believing. Subliminal perception and other nonconscious mental states are cases of its seeming to one that things are a particular way without its also seeming to one that things seem to be that way. They are cases of being aware *of* something even though the psychological state in virtue of which one is aware of that thing is not a conscious state. Perceiving and believing couldn't fail to be conscious if there were no distinction between seeming and seeming to seem.<sup>11</sup>

Given Dennett's rejection of a distinction between seeming and seeming to seem, it's no surprise that he sometimes explicitly elides the distinction between one's being conscious of something and a mental state's being conscious. FPO, he tells us, "denies the possibility in principle of consciousness of a stimulus in the absence of the subject's belief in that consciousness" (1991a, p. 132). But adoption of the intentional stance will sometimes enable one to discern that the subject is conscious of something even if the subject sincerely denies being conscious of it. Dennett's denial that one can be conscious of something without realizing that one is not only rules out subliminal perception; it also encounters difficulty with possible conflicts between third-person ascriptions by the intentional stance of

---

11. Dretske's account of mental states' being conscious runs afoul of the same difficulty. On his view, a state is conscious if in virtue of being in that state one is conscious of something; "experiences and beliefs are conscious, not because you are conscious of them, but because, so to speak, you are conscious *with* them" (1995, pp. 280–281). But nonconscious states make one aware of things no less than conscious states, just not consciously aware.

another person's mental states and the FPO insistence that people's beliefs about their own mental states is the last word about those states.

Dennett of course recognizes that subliminal perception and unconscious believing occur. But he holds it's a mistake to describe these things in ordinary folk-psychological terms. Rather, he insists, we should describe these occurrences in terms of what he calls "events of content fixation" (1991a, pp. 365, 457–458). Thinking of these cases as genuine perceiving and believing, thereby applying the categories of folk psychology to them, fails to recognize their fleeting nature and the rough-and-tumble competition among them to precipitate into stable, well-ordered, conscious folk-psychological states.

Dennett's claim that nonconscious states are mere events of content fixation, and not genuine folk-psychological states, is of a piece with his treatment of the puzzles about timing. When the long-haired woman without eyeglasses runs by, events of content fixation that result from that visual input compete with events of content fixation that stem from the memory of one's friend with short hair and eyeglasses. The competition happens to result in a conscious perception of the short-haired friend with eyeglasses. And if there is no fact of the matter about whether the memory intrudes before or after the visual input makes it to consciousness, there are no folk-psychological states, properly so called, prior to one's conscious perception, only competing events of content fixation.

But this doesn't give the relevant nonconscious states their due. We taxonomize states in folk-psychological terms when they have mental content, even if they are fleeting and in competition for ultimate standing in our stream of consciousness. And the visual input of one woman and memory of another have the right kind of content, however fleeting and unstable the states may be. And folk-psychological states can fail to be conscious and still be long lasting and have a stable effect on our mental lives. There is no non-question-begging reason to deny full folk-psychological standing to contentful states one is unaware of being in.

If one recasts issues about mental states generally in terms of neural function, as Dennett sometimes does (e.g., 2015), then it may well seem inviting to dismiss states that aren't conscious as merely neural, not appropriately taxonomized in terms of folk-psychological categories. But the properties that matter to our folk-psychological taxonomizing are those that pertain to perceptible properties, such as color and sound, and intentional content; consciousness isn't needed for states to have those properties.

Our problem is to square the first- and third-person perspectives on the mental, in a way that does justice to both. Dennett's approach wisely starts from the third person, recognizing that first-person resources are too sparse to allow us to construct a satisfactory account of the third-person aspect of mind. Nonetheless, his approach arguably founders on a set of closely connected

issues. On FPO one's beliefs about one's psychological states are the last word about what mental states one is in. And Dennett rejects taxonomizing any nonconscious states as mental. So such states cannot occur without being conscious; states that aren't conscious cannot be genuine psychological states. States of seeming are folk-psychological states; so there are no states of seeming that aren't conscious. And since no states of seeming occur without being conscious, there is no room for a distinction between how things seem and how it seems to one that they seem.

The way FPO construes consciousness and the first-person aspect of mind accordingly makes for a difficulty in having them square with the third-person aspect of mind. If we accord a robust privilege to first-person beliefs about what one is conscious of, there are only two ways for our first- and third-person perspectives on psychological phenomena to fit together. One way, unavailable to Dennett, would be to downgrade our third-person access by denying that we often can tell what psychological states others are in. That is not only unrealistic; it also flies in the face of Dennett's sensible realism about the states we discern by the intentional stance.

The other way to square things is to downgrade the nature of the states discernible by the intentional stance when the individual in question is unaware of being in these states. That's Dennett's way; we mustn't taxonomize such states in folk-psychological terms, but should regard them instead as mere events of content fixation. But there is every reason to apply ordinary folk-psychological categories to such states. Given FPO, there seems no third way to have first and third person fit together. Dennett's account does not, after all, provide a successful path from the third-person aspect of mind to its first-person aspect.

#### 4. The Higher-Order Alternative

The most promising way to get an account on which first- and third-person aspects mind fit comfortably together is to start from the third person. But it's important to bear in mind, as noted in section 1, that we can do justice to intuitions that involve the first-person perspective without counting them all as true. Not all appearances in any area are accurate. When they aren't, however, it's seldom appropriate simply to dismiss them; we must explain why it is that they strike us as compelling. Even when we can't save the intuitive appearances because they aren't accurate, explaining why we have those appearances is a sound way to do justice to them.

The first-person appearance that has been causing trouble is the idea that first-person access is infallible, since that more than anything else blocks comfortable coexistence with any third-person approach to mental states. If the strategy of

starting with the third person and constructing a satisfactory first-person picture of mind must accommodate infallibility, it will almost certainly fail. By contrast, if first-person access is not infallible, it's open for one's subjective sense of what mental state one is in sometimes to be inaccurate; one may on occasion be aware of oneself as being in states of seeming that differ from the states of seeming one is actually in. It would then seem to one that one's states of seeming are different from what they actually are.

Dennett's FPO, by preserving what is in effect an operationalist version of infallibility, won't give us what we need. But we can get that with an account that is a close cousin of Dennett's. On that account, what it is for a mental state to be conscious consists in one's being aware of oneself, in the right way, as being in that state; a state is conscious if one has a suitable higher-order awareness that one is in that state. This is close to Dennett's view, since it accommodates his idea that the consciousness of mental states consists in what is given "to be taken one way rather than another."

But the alternative, higher-order account denies the operationalist way that FPO spells out what it is for the given to be taken. Being taken is not a matter of "writing it down" in memory" or "of the effects of various vehicles of content on subsequent action (and hence, of course, on memory)" (1991a, p. 132). Rather, it is simply a matter of one's being aware of oneself as being in the mental state in question.<sup>12</sup> FPO insists that "there are no fixed facts about the stream of consciousness independent of particular probes" (1991a, p. 138; cf. p. 275). The alternative account takes probes as providing symptoms or expressions, verbal or otherwise, of the higher-order awareness in virtue of which mental states are conscious.

Such an account is compelling for independent reasons. Once we acknowledge that genuine folk-psychological states sometimes occur without being conscious, it's hard to avoid some type of higher-order theory about what it is for such states to be conscious. No mental state is conscious if one is wholly unaware of being in it. So if there is solid evidence that somebody thinks or sees something but the person sincerely denies doing so, the thinking or seeing simply isn't conscious. And that's equivalent to a state's being conscious only if one is aware of being in it.

Not any type of higher-order awareness will do; if one is aware of oneself as thinking or seeing something only because of applying a theory to oneself, experimental findings, observing one's own behavior, or taking somebody else's word

---

12. That higher-order awareness will typically, perhaps always, occur very slightly after the onset of the state it makes one aware of being in. But there is no reason to think that the process involves memory of any sort. I am grateful to Mandik for pressing this question.

for it, such thinking or seeing will not be conscious. One must be aware of oneself as being in the state in some way that is independent of conscious observation and inference, that is, observation and inference of which one is aware. It must seem subjectively to one that one's awareness of the state does not rely on inference or observation.<sup>13</sup>

The traditional view is that first-person access is direct and unmediated. This goes well beyond what we have any reason to believe. Our first-person access, based as it is on the relevant state's being conscious, seems subjectively to involve no mediation by observation or inference. That doesn't show that no mediation occurs, only that if there is any it isn't conscious; we are unaware of any mediation that does occur. These considerations do justice to the first-person impression of actual lack of mediation not by crediting it with being true, but by explaining why it seems compelling.

Subjective impressions about lack of mediation aside, we can also explain why it's inviting to regard first-person access as infallible, again without taking that traditional doctrine to be true. There is no subjective check on the higher-order awareness we have of our conscious states. And because there is no subjective appeal beyond our higher-order awareness, we have a subjective impression that such awareness is the final word about our mental lives. But it is only the final subjective word about our mental states, the final word simply about how we are aware of the states, not about what those states actually are.

There is more to be said about the kind of higher-order awareness that figures in a mental state's being conscious. I have argued elsewhere (e.g., 1986, 2005) that we can best understand that higher-order awareness as consisting in one's having a thought to the effect that one is in the state in question. Having such higher-order thoughts (HOTs) provides a more satisfactory explanation than any other type of higher-order awareness.

One especially important advantage of construing one's awareness of each conscious state as a HOT to the effect that one is in that state is that it offers a ready explanation of why, barring special deficits or unusual circumstances, psychological states are verbally reportable just in case they're conscious. Thoughts are expressible in speech; if I have a thought that it's raining, I can express that thought by saying that it's raining. Similarly, if one has a thought that one is in a particular mental state, one can express that thought by saying that one is in that

---

13. One's having consciously inferential or observational reason to be aware of oneself as being in a state does not prevent the state from being conscious so long as one is also aware of oneself as being in the state in a way that subjectively seems independent of inference and observation, contrary to an objection of Block's (2011, p. 446).

state. Expressing a HOT simply is reporting the state that the HOT is about. The coincidence of a state's being reportable and its being conscious barring special circumstances is best explained on the hypothesis that the way we're aware of a conscious state is by having a thought that one is in that state. No other type of higher-order awareness explains the ready reportability of states that are conscious.

HOTs need not be conscious, and indeed seldom are; they're conscious only if there's a third-order thought about the second-order thought. Conscious HOTs likely figure in introspective consciousness. When we introspect, we are not just aware of being in a mental state; we are also aware that we are thus aware. We attentively focus on our awareness of the introspected state; we are accordingly consciously aware of that state.<sup>14</sup>

Ordinary, nonconscious HOTs, by contrast, explain in virtue of what a conscious state differs from a mental state that isn't conscious. Since a mental state fails to be conscious only if one is wholly unaware of being in it, being aware of it in some way is necessary for the state to be conscious. Subjectively noninferential reportability points to HOTs as the way one is aware of one's conscious states. But as long as the HOT isn't itself conscious, one is aware of the state but not consciously aware of it; so there is no introspective awareness and no conscious judgment that one is in the state.<sup>15</sup>

Acknowledging genuinely psychological states that aren't conscious allows a good fit between first and third person. We have pretty reliable access to others' psychological states, essentially by the kind of tools Dennett describes in connection with the intentional stance. When we have solid third-person reason to ascribe a psychological state to somebody that the person sincerely denies being in, we typically posit a psychological state that isn't conscious but is still taxonomized in terms of standard folk-psychological categories.

Such positing of nonconscious states will rely simply on folk-psychological considerations; novels and plays have been replete with examples long before the influence of Freud. There is no need to demote such nonconscious states to the status of mere events of content fixation. On a higher-order theory, such a state

---

14. Amy Kind (unpublished manuscript) has forcefully argued that such attentive focus is not present in all cases of what are generally regarded as introspectively conscious states. Still, many introspectively conscious states do involve some deliberate, attentive focus, and an account that appeals to third-order thoughts—that is, to our being aware of being aware of our introspected states—arguably does apply to them.

15. Contrary to Block's (1995, p. 235) apparent assimilation of consciousness that involves HOTs to monitoring or reflective consciousness, by which he evidently has in mind introspective consciousness. Such assimilation likely rests on a tacit assumption that HOTs are invariably conscious.



is genuinely folk-psychological despite its being unaccompanied by a suitable higher-order awareness.

FPO “brusquely denies the possibility in principle of consciousness of a stimulus in the absence of the subject’s belief in that consciousness” (1991a, p. 132). That denial concerns one’s being conscious of observable objects and processes; it does not on that account bear on what it is for a psychological state to be a conscious state.

The consciousness of mental states is a matter of mental appearance; it is the way our mental lives appear to us. We are likely in a great many states of which we are wholly unaware but, because of their content, are appropriately described in folk-psychological terms. Mental states are conscious only if we appear subjectively to be in them. And because these states have content, they reflect the way things appear to us. So there is a second level of appearing when the states are conscious, the subjective appearance that one is in those states. That is what seeming to seem consists in.

But appearance can be illusory. Just as a first-order appearance of a red square can be illusory, so can the higher-order, subjective appearance that one is seeing a red square. If there is higher-order awareness, it must be that there can also be higher-order misrepresentation. This has seemed to many to be the undoing of any higher-order theory. But there is no reason to doubt that consciousness can misrepresent what mental states one is in (Weisberg, 2008, 2011a, b; Rosenthal, 2011, 2012).<sup>16</sup> Indeed, consciousness misrepresents whenever one is in a psychological state that isn’t conscious, at least insofar as we typically assume if only tacitly that consciousness reveals all our current mental states. So consciousness in effect represents one as not being in any mental state that isn’t conscious. And we are routinely aware of our color sensations in a relatively generic way, for example, simply as generic sensations of red, though the visual sensations doubtless reflect a far more specific shade. That, too, is a type of higher-order misrepresentation, albeit mild and innocuous.

Higher-order misrepresentation also occurs, however, in more dramatic ways. A striking example occurs in a particular type of change blindness in which the display changes during a saccade, when no significant retinal information reaches visual cortex. Using eye trackers to time the changes, John Grimes (1996) found a significant number of subjects who failed to notice even highly salient changes, for example, a change of color from green to red

---

16. And as Weisberg (2008, 2011a, b) forcefully shows, it would be no obstacle to misrepresentation by consciousness for the higher-order awareness to be intrinsic to the state it makes one aware of.



in a parrot that occurs centrally in the display and occupies about 25% of it.<sup>17</sup>

After the parrot changes from green to red, retinal information of red presumably reaches visual cortex; one has a sensation of red. But if one notices no change, what it's like for one is that one is still consciously seeing green; one remains aware of oneself as having a sensation of green. One's higher-order awareness misrepresents one's first-order qualitative mental state. Doubtless random saccades often result in this kind of effect in everyday life.

Michael A. Cohen and Dennett have recently argued against any view that dissociates consciousness from all cognitive function. They initially describe such cognitive function very inclusively, as including "attention, working memory, language, decision making, motivation etc." But they immediately go on to describe it more narrowly, simply as "verbal report, button pressing etc." (2011, p. 358). And they note in a follow up to their original article that the cognitive functions they focus on "are all the products of cognitive access" (2012, p. 140).

Cohen and Dennett stress that they are not identifying a mental state's being conscious with its having some such cognitive function, but rather simply denying that one can detect a conscious state in others independent of any such function (2012). And that appeal to cognitive function is in keeping with the importance of the third-person considerations urged in section 1. Still, focusing on the broad list of cognitive functions Cohen and Dennett initially offer risks distracting from what it is in virtue of which mental states occur consciously.

Indeed, all the cognitive functions on the more inclusive list except sincere verbal report can arguably occur in connection with psychological states that aren't conscious (Rosenthal, 2008; 2012, section 5). Even attention, which Cohen and Dennett highlight (2011, pp. 359–360), demonstrably occurs in the absence of conscious awareness (e.g., Norman, Heywood, & Kentridge, 2013, 2015; van Boxtel, Tsuchiya, & Koch, 2010; see Montemayor & Haladjian, 2015, for a useful review). So not only must we avoid identifying those functions with consciousness; most can't reliably serve even as reliable indicators of a mental state's being conscious.

Cohen and Dennett's shorter list, "verbal reports, button presses," does reflect what it is for a mental state to be conscious, and singles out the pivotal indicator. A sincere denial of being in a mental state shows that if, despite that denial, the individual is actually in the state, that state is not conscious. Sincere denial is decisive because it reflects lack of any awareness that one is in the state. So a sincere

---

17. Thanks to Dennett for having first alerted me many years ago to this work.

report is similarly decisive that the reported state is conscious.<sup>18</sup> And since sincere reports express one's awareness of what one reports, such reports point to higher-order awareness of a state as constitutive of its being conscious.

Indeed, this is the upshot of the "perfect experiment" Cohen and Dennett (2011, pp. 361–362) describe in support of their appeal to cognitive function. In that thought experiment we imagine the cortical area responsible for conscious color experiences having been severed surgically from every other cortical area.<sup>19</sup> Because the area is detached, the patient has no access to whatever occurs there, and so would sincerely deny having any color experiences. Cohen and Dennett conclude that so far as consciousness is concerned it doesn't matter what happens in the detached area; sincere report and denial are the last word about consciousness. And though they don't go on to say as much, the best explanation of why sincere report is decisive about mental states' being conscious is that it expresses one's higher-order awareness of those states, and that that's constitutive of the states' being conscious.

Cohen and Dennett's main target is the view advanced by Ned Block (2005, 2007) that what Block calls phenomenal consciousness is wholly independent of and can occur in the absence of any form of cognitive access. That would also undermine any higher-order theory, since higher-order awareness is a type of cognitive access. Block proposes identifying phenomenal consciousness with particular neural correlates; we could then determine its presence neurally. But as Cohen and Dennett argue, one cannot establish neural correlates of any type of consciousness without an independent way to tell when such consciousness occurs.<sup>20</sup> Any such independent way will inevitably be psychological, and so detectable in others only by their behavior; hence the crucial role of the cognitive functions Cohen and Dennett appeal to.

Dennett would resist inferring from cognitive function to higher-order awareness; we should not, he would urge, construe the cognitive access he and Cohen insist on as a form of higher-order awareness. One way he might resist that construal would be by appeal to the intentional stance as the arbiter of objective psychological reality. The intentional stance underwrites ascribing first-order

---

18. At least if the report does not rely on conscious inference or observation. The appeal to sincerity does not go beyond what is intersubjectively accessible; we can often tell, even if not perfectly, when others' speech acts are sincere.

19. Cohen and Dennett acknowledge that the thought experiment is highly unrealistic (2011, p. 361).

20. Indeed, Block (e.g., 1995, 2005, 2007) arguably recognizes the need in practice to rely on signs of cognitive access to determine the presence of phenomenally conscious states in others, despite his insistence that phenomenal consciousness is independent of any such access (see esp. 2007, p. 487, noted by Cohen & Dennett, 2011, p. 360).

psychological states, but does not also sustain ascribing higher-order awareness of those states.

But the intentional stance is also neutral about whether the states ascribed to others are conscious; it tells us what psychological states others are in, but not whether they're conscious. Ascription by way of the intentional stance rests on a background assumption of an individual's rationality, and rationality is independent of whether the relevant psychological states in question are conscious. If believing that *p* and desiring *a* makes it rational to do a particular thing, that's independent of whether the belief and desire are conscious (Rosenthal, 2008, section 2; 2012, section 5). So the failure of the intentional stance to ascribe higher-order awareness can't tell against explaining a mental state's being conscious in terms of higher-order awareness, since the intentional stance is also silent about whether psychological states are ever conscious. Indeed, the intentional stance by itself doesn't even sustain ascribing cognitive access of the sort Cohen and Dennett acknowledge as crucial to the study of consciousness.

Even if the intentional stance is normally neutral about whether an ascribed state is conscious, one might urge that it need not always be.<sup>21</sup> If somebody says, "I think it's raining," that speech act, taken literally, reports one's thought that it's raining, and it's natural to take that report to be independent of conscious inference or observation. So the speech act would express a HOT that one thinks it's raining, and we should accordingly ascribe to the speaker a conscious thought that it's raining.

But no appeal to explicit reports of one's mental states plays any role whatever in the intentional stance; rather, such an appeal would supplement the intentional stance. The intentional stance is concerned solely with which thoughts and desires we should posit, in a folk-psychological way, to make an individual's speech and other behavior rational. And the states that would make it rational to say "It's raining" are exactly the same as those that would make it rational to say "I think it's raining"; performing either speech act is wholly rational if one has a thought that it's raining, conscious or not.<sup>22</sup>

The exclusive reliance by the intentional on rationality results in its treating the two types of speech act as equivalent. But they clearly are not, since they differ in truth conditions; each can be true when the other is not. This echoes

---

21. I am grateful to David Pereplyotchik for raising this possibility.

22. Verbally expressed thoughts are typically conscious, but their being conscious contributes nothing to the rationality of the speech acts that express them. Indeed, the only available account of why verbally expressed thoughts are conscious that doesn't simply stipulate an unexplained tie between consciousness and language makes no appeal to rationality. See Rosenthal (2005, chapter 10).

Wittgenstein's focus on use to the exclusion of truth conditions, which also leads to one's seeing as equivalent the speech acts of saying that *p* and saying that one thinks that *p* (Wittgenstein, 1953, I, x).

It's inviting to speculate that this limitation on the intentional stance leads Dennett to his firm rejection of explaining consciousness by appeal to higher-order awareness (Dennett, 1991a, chapter 10, esp. pp. 314–320). If the intentional stance is the last word about objective psychological reality but can't distinguish conscious from nonconscious psychological states, then if any psychological states are conscious, all of them are. We must then demote any nonconscious states that function much as conscious psychological states do to the status of mere events of content fixation, and not countenance them as psychological states at all.

But though the intentional stance is blind to the difference between conscious and nonconscious psychological states, we can supplement that stance with the sincere reports and button presses that Cohen and Dennett appeal to, which serve as evidence of the cognitive access Dennett accepts as pivotal to the study of consciousness. Doing so preserves the reliance on third-person considerations, but also provides the resources needed for an appeal to higher-order awareness in explaining consciousness, and with it a clear and robust distinction between seeming and seeming to seem.

## Works Cited

- Alter, T., & Walter, S. (2007). (Eds.). *Phenomenal concepts and phenomenal knowledge: New essays on consciousness and physicalism*. Oxford, UK: Oxford University Press.
- Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18, 227–247.
- Block, N. (2005). Two neural correlates of consciousness. *Trends in Cognitive Sciences*, 9, 46–52.
- Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences*, 30, 481–499.
- Block, N. (2011). Response to Rosenthal and Weisberg. *Analysis*, 71, 443–448.
- Breitmeyer, B., & Ögmen, H. (2006). *Visual masking: Time slices through conscious and unconscious vision*. (2nd ed.). New York, NY: Oxford University Press.
- Cohen, M., & Dennett, D. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, 15, 358–364.
- Cohen, M., & D. Dennett, (2012). Response to Fahrenfort and Lamme: Defining reportability, accessibility and sufficiency in conscious awareness. *Trends in Cognitive Sciences*, 16, 139–140.
- Dennett, D. C. (1978). *Brainstorms*. Cambridge, MA: MIT Press.
- Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

- Dennett, D. C. (1991a). *Consciousness explained*. Boston, MA: Little, Brown.
- Dennett, D. C. (1991b). Real patterns. *Journal of Philosophy*, 88, 27–51; reprinted in Dennett (1998), pp. 95–120.
- Dennett, D. C. (1996). *Kinds of minds: Toward an understanding of consciousness*. New York, NY: Basic Books.
- Dennett, D. C. (1998). *Brainchildren: Essays on designing minds*. Cambridge, MA: MIT Press.
- Dennett, D. C. (2005). *Sweet dreams: Philosophical obstacles to a science of consciousness*. Cambridge, MA: MIT Press.
- Dennett, D. C. (2009). The part of cognitive science that is philosophy. *Topics in Cognitive Science*, 1, 231–236.
- Dennett, D. C. (2013). *Intuition pumps and other tools for thinking*. New York, NY: Norton.
- Dennett, D. C. (2015). Why and how does consciousness seem the way it seems? In T. Metzinger & J. M. Windt (Eds.), *Open MIND* (pp. 387–398). Frankfurt am Main, Germany: MIND Group. Retrieved from <http://open-mind.net/collection.pdf>.
- Dennett, D. C., & Kinsbourne, M. (1992). Time and the observer: The where and when of consciousness in the brain. *Behavioral and Brain Sciences*, 15, 183–201.
- de Gelder, B., Vroomen, J., Pourtois, G., & Weiskrantz, L. (1999). Non-conscious recognition of affect in the absence of striate cortex. *NeuroReport*, 10, 3759–3763.
- Dretske, F. (2000). *Knowledge, perception, and belief: Selected essays*. Cambridge, England: Cambridge University Press.
- Fernandez-Duque, D., & Thornton, I. M. (2000). Change detection without awareness: Do explicit reports underestimate the representation of change in the visual system? *Visual Cognition*, 7, 324–344.
- Grimes, J. (1996). On the failure to detect changes in scenes across saccades. In K. Akins (Ed.), *Perception* (pp. 89–110). New York, NY: Oxford University Press.
- Kind, A. *Thin Introspection*. (unpublished manuscript).
- Kolers, Paul A., & von Grünau, M. (1976). Shape and color in apparent motion. *Vision Research*, 16(4), 329–438.
- Koyama, T., McHaffie, J., Laurienti, P., & Coghill, R. (2005). The subjective experience of pain: Where expectations become reality. *Proceedings of the National Academy of Science*, 102, 12950–12955.
- Kripke, S. (1980). *Naming and necessity*. Cambridge, MA: Harvard University Press.
- Laloyaux, C., Destrebecqz, A., & Cleeremans, A. (2003). Implicit change identification: A replication of Fernandez-Duque and Thornton. *Journal of Experimental Psychology: Human Perception and Performance*, 32, 1366–1379.
- Montemayor, C., & Haladjian, H. (2015). *Consciousness, attention, and conscious attention*. Cambridge, MA: MIT Press.
- Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, 83, 435–450.

- Norman, L., Heywood, C., & Kentridge, R. (2013). Object-based attention without awareness. *Psychological Science*, 24, 837–843.
- Norman, L., Heywood, C., & Kentridge, R. (2015). Exogenous attention to unseen objects? *Consciousness and Cognition*, 35, 319–329.
- Ögmen, H., & Breitmeyer, B. (2006). (Eds.). *The first half second: The microgenesis and temporal dynamics of unconscious and conscious visual processes*. Cambridge, MA: MIT Press.
- Quine, W. V. (1960). *Word and object*. Cambridge, MA: MIT Press.
- Quine, W. V. (1985). States of mind. *Journal of Philosophy*, 82, 5–8.
- Rosenthal, D. (1986). Two concepts of consciousness. *Philosophical Studies*, 49, 329–359.
- Rosenthal, D. (1989). Philosophy and its history. In A. Cohen & M. Dascal (Eds.), *The institution of philosophy* (pp. 141–176). Peru, IL: Open Court, 1989.
- Rosenthal, D. (2005). *Consciousness and mind*. Oxford, UK: Clarendon Press.
- Rosenthal, D. (2008). Consciousness and its function. *Neuropsychologia*, 46, 829–840.
- Rosenthal, D. (2009). Philosophy and its teaching. In R. Talisse & M. Eckert (Eds.), *A teacher's life: Essays for Steven Cahn* (pp. 67–83). Lanham, MD: Lexington.
- Rosenthal, D. (2010). How to think about mental qualities. *Philosophical Issues*, 20, 368–393.
- Rosenthal, D. (2011). Exaggerated reports: Reply to Block. *Analysis*, 71, 431–437.
- Rosenthal, D. (2012). Higher-order awareness, misrepresentation, and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367, 1424–1438.
- Sellars, W. (1956). Empiricism and the philosophy of mind. In H. Feigl & M. Scriven (Eds.), *Minnesota studies in the philosophy of science* (Vol. 1, pp. 253–329). Minneapolis: University of Minnesota Press. (Reprinted in *Science, perception and reality*, by W. Sellars, London: Routledge and Kegan Paul, 1963, and Atascadero, CA: Ridgeview, 1991, pp. 127–196).
- Sellars, W. (1968). *Science and metaphysics*. London: Routledge and Kegan Paul. (Reprinted by Ridgeview, 1992, Atascadero, CA).
- Shoemaker, S. (2003). The inverted spectrum. In S. Shoemaker, *Identity, cause, and mind: Philosophical essays* (2nd ed., pp. 327–357). Oxford, UK: Clarendon.
- Strawson, P. F. (1959). *Individuals: An essay in descriptive metaphysics*. London: Routledge.
- van Boxtel, J. A., Tsuchiya, N., & Koch, C. (2010, December 20). Consciousness and attention: On sufficiency and necessity. *Frontiers in Psychology* [Review article]. Retrieved from <http://dx.doi.org/10.3389/fpsyg.2010.00217>
- Weiskrantz, L. (1997). *Consciousness lost and found: A neuropsychological exploration*. Oxford, UK: Oxford University Press.
- Weisberg, J. (2008). Same old, same old: The same-order representation theory of consciousness and the division of phenomenal labor. *Synthese*, 160, 161–181.

- Weisberg, J. (2011a). Abusing the notion of what-it's-like-ness: A response to Block. *Analysis*, 71, 438–443.
- Weisberg, J. (2011b). Misrepresenting consciousness. *Philosophical Studies*, 154, 409–433.
- Wittgenstein, L. (1953). *Philosophical Investigations* (G. E. M. Anscombe, Trans.) Oxford, UK: Basil Blackwell.

## 5.2 REFLECTIONS ON DAVID ROSENTHAL

Daniel C. Dennett

This masterful essay, a model of clarity, objectivity, and constructive thinking, exposes a tension in my discussions of *seeming*. The preamble section sets out the all too familiar conflict between first-person and third-person views in ways that were illuminating to me—someone who has been thinking about these very points for half a century and more. In particular, it provides the background for Rosenthal’s explanation of how I have used the intentional stance to build an account of consciousness from the third-person perspective that attempts to do justice to the richness of our first-person intuitions while at the same time showing why our intuition that we have incorrigible access to the contents of our minds is compelling but mistaken. He points out that whereas I have echoed Nagel’s insistence that there is no distinction in consciousness between seeming and reality, my reasoning is not Nagel’s: where Nagel indulges in mystery, I attempt to account for this with first-person operationalism (FPO), which has a decidedly third-person, naturalistic basis. In my account, the vaunted incorrigibility of the first-person point of view turns out to be (trivially) *constituted by* our convictions about how it seems, in much the same way that a center of gravity is incorrigibly located in any object at the point which, if supported, would permit the object to remain in equilibrium in any position. Don’t ask how we know this remarkable juxtaposition; it’s in effect true by definition. “We aren’t wrong about our conscious experiences because when it comes to conscious experiences there is nothing beyond our beliefs to be wrong about” (p. 144).

Rosenthal’s reconstruction of my argument about Orwellian and Stalinesque phenomena, and the question of whether there is always a fact of the matter is deft and fair, and it leads him to an observation about the intentional stance that has gone all but unmarked: the



intentional stance doesn't in itself distinguish between conscious beliefs and unconscious beliefs:

Your observations can reveal the states I'm objectively in, whereas my beliefs about what conscious states I'm in is the last word only about how my subjective stream of consciousness appears to me. How my stream of consciousness appears to me is itself objective; it appears to me in one way and not another, and my beliefs about that are the last word about how it appears to me. But that objective matter is different from the objective matter your observations of real patterns reveal, namely, beliefs and desires I have independently of how my mental life appears to me. (p. 148)

How, then, can I “emphatically reject” (as I did) a distinction between how it seems to me and how it seems to seem to me? Rosenthal has in fact uncovered a large and embarrassing contradiction in my 1991 position, which I unwittingly papered over in my campaign to shake people's allegiance to *qualia*. Unpacking this confusion of mine yields some nice surprises—at least to me.

First, let me grant Rosenthal his central point: my denial that there is a distinction between seeming and seeming to seem is belied by phenomena such as “subliminal perception and other nonconscious mental states” (p. 151). Among such phenomena are standard cases of self-deception, in which, for instance, I sure seem to the onlookers to distrust my neighbor while seeming to myself, as I avow sincerely, to trust him with my life. Others are less common cases of more symmetrical confusion. In fact, I presented a particularly vivid case of this phenomenon in *Consciousness Explained* (1991; hereafter, CE) without drawing attention to—without noticing—this conflict: my misadventure as a baseball umpire calling a crucial play at first base (p. 248). How could I not see that this was a case of seeming to seem, fitting Rosenthal's analysis perfectly?

It has taken me some considerable reflection to solve this mystery of self-interpretation. My blindness grew out of my campaign to deny a similar but logically independent presupposition that is also common among the lovers of *qualia*: the compelling conviction that *really seeming* to see a purple cow somehow involves purple *qualia* as *real* properties of my experience, in contrast with, say, a robot seeming to see a red, white and blue American flag when confronted with a complementary color afterimage. A robot equipped with a good model of human color vision might superficially seem to see a red white and blue flag, according to this popular view, but this wouldn't be real seeming, but at best some “merely behavioristic” kind of seeming.

Rosenthal's version of the distinction has nothing to do with *qualia*. It is about how there can be a conflict between what I seem (to other interpreters) to believe,

etc., and what I seem (to myself) to believe; when such a conflict arises, I can seem to judge that the runner is out (just look at my vigorous hand signal) while also seeming (or seeming to seem) to judge the runner to be safe (just listen to what I say). It is sometimes said that this sense of “seem” is a mere indicator of epistemic modesty (“Is that a deer running through the woods, or does it just seem to me to be one?”), not the assertion about any *real* seeming (“Oh wow, I seem to see an undulating checkerboard with Day-Glo green and silver squares!”). It is this idea of real seeming over and above the non-committal merely epistemic seeming that I still wish to deny. (See Andy Clark’s chapter, 7.1, and my Reflections on him, 7.2, for some new slants.)

In short, I am not recanting my dismissal of qualia, as philosophers consider them, but I am taking on board Rosenthal’s claim that there are plenty of cases of where the third-person use of the intentional stance conflicts with the first-person use. Indeed, it is in order to handle these possibilities gracefully that heterophenomenology declares subjects to be logically incorrigible about what they *seem* to experience. That is FPO in action, letting subjects constitute their heterophenomenological worlds by fiat in their considered judgments about what it is like to be them, a preliminary tactic that fixes *what is in need of explanation* while leaving wide open the prospect of discovering that subjects are, in spite of the confidence of their convictions, wrong about what is really going on in them. Thus Roger Shepard and Zenon Pylyshyn can agree that their subjects seem to be rotating images—that’s what they all insist on, if you ask them—while disagreeing about what is really going on inside them. (For an early version of my FPO applied to this controversy, see my “Two Approaches to Mental Images” in *Brainstorms*, 1978, where I distinguished the “ $\beta$ -manifold” of intentional objects from the underlying machinery.)

But now, what of Rosenthal’s further conclusion, that my FPO fails in its purpose? “Dennett’s account does not, after all, provide a successful path from the third-person aspect of mind to its first-person aspect” (p. 153). I am not persuaded of this, since I think I can happily acknowledge that the distinction Rosenthal draws between seeming and seeming to seem is honored by my account after all, but doing so requires me to remove another bit of papering over, which (it seems to me) is also a problem for Rosenthal’s “close cousin” to my view, the latest version of his famous HOT (Higher Order Thought) theory of consciousness: “what it is for a mental state to be conscious consists in one’s being aware of oneself, in the right way, as being in that state; a state is conscious if one has a suitable high-order awareness that one is in that state” (p. 154).

I have always been drawn to some aspects of HOT theory, as I explained in *CE*. In particular, HOT theory provides a sensitive account of one very tricky feature of human consciousness: reportability in language. There is a big difference

between *expressing* a mental state and *reporting* a mental state; the poker player who can't manage to maintain a good—uninformative—poker face involuntarily exhibits facial expressions, hand gestures, postures nicely known as “tells” in the world of poker. Discerning the tells of the other players as tantamount to “reading” their minds, a tremendous advantage to the observant player. In spite of the words “tell” and “read,” this is not verbal communication; tells are instances of negligent self-exposure, not intended speech acts. Tells express mental states without reporting mental states. But when a person does *report* a mental state by *saying* (in words, or via prearranged button press) “the circle on the left seems to be larger,” this intentional action ipso facto *expresses* a mental state, the higher-order belief or thought that the circle on the left seems to be larger, a claim *about* a mental state.

This is all very well when we are considering the heterophenomenology of human consciousness, but what about animal consciousness? In spite of the ubiquitous but misleading practice of psychologists who let themselves speak of their animal subjects “telling” them which state they are in by performing one highly trained action or another (usually an eye blink or button press) this *misattributes* a communicative intention to the subject. Monkeys and rats are not trained to *communicate* with the experimenters (the way human introspectors were trained by the various early schools of introspectionism); they are trained to *inform* the experimenters in much the way a poker player might be patiently and subliminally encouraged by another player to exhibit tells. The animal subjects, unlike human subjects in heterophenomenological experiments, do not have to know what they are doing. But then it follows that the states of mind revealed by such tells are importantly *unlike* the states of mind we humans spend so much time and ingenuity conferring about.

I have long stressed the fact that human consciousness is vastly different from the consciousness of any other species, such as apes, dolphins, and dogs, and this “human exceptionalism” has been met with little favor by my fellow consciousness theorists. Yes, of course, human beings, thanks to language, can do all sorts of things with their consciousness that their language-less cousin species cannot, but still, goes the common complaint, I have pushed my claims into extreme versions that are objectionable, and even offensive. Not wanting to stir up more resistance than necessary to my view, I have on occasion strategically soft-pedaled my claims, allowing animals to be heterophenomenological subjects (of sorts) thanks to their capacity to *inform* experimenters (if not *tell* them), but now, my thinking clarified by Rosenthal's, I want to recant that boundary blurring and re-emphasize the differences, which I think Rosenthal may underestimate as well. “Thoughts are expressible in speech,” he writes (p. 155), but what about the higher-order thoughts of conscious animals? Are they? They are not *expressed* in speech, and

I submit that it is a kind of wishful thinking to fill the minds of our dogs with thoughts of that sophistication. So I express my gratitude to Rosenthal for his clarifying account by paying him back with a challenge: how would he establish that non-speaking animals have higher-order thoughts worthy of the name? Or does he agree with me that the *anchoring* concept of consciousness, human consciousness, is hugely richer than animal consciousness on just this dimension?

Rosenthal says his view is close to mine, since it accommodates my idea that the consciousness of mental states consists in what is given “to be taken in one way rather than another.” But he denies

the operationalist way that FPO spells out what it is for the given to be taken. Being taken is not a matter of ‘writing it down’ in memory or ‘of the effects of various vehicles of content on subsequent action (and hence, of course, on memory)’ (1991a, p. 132) Rather, it is simply a matter of one’s being aware of oneself as being in the mental state in question. (p. 154)

Simply? And then what happens? Or: and then what *may* happen? The agent who is thus aware of being in that mental state is, in virtue of that awareness, able to hinge almost any action on that mental state. (Debner & Jacoby, 1994. For discussion see Dennett 2017, *BBB*) But do we have any grounds for generalizing the experimental animal’s telltale blink or button-press to a more general ability to *use what it knows about what mental state it is in*? It is a bit of a stretch to call the animal’s response the expression of a *belief*. If we want to secure some form of higher-order thought, we will have to find many ways of assaying the versatility of this ability. To date, the only research I know that makes an inroad on this issue are the metacognition experiments that apparently show animals can use an assessment of their confidence to choose between a risky high-payoff task and an easy low-payoff task (e.g., Shields, Smith, Guttmanova, & Washburn, 2005).

Let me respond, finally, to two footnotes in Rosenthal’s essay. A casual reading of footnote 8 might conclude that I coined “intuition pump” in (1991), in *CE*, but in fact, it was much earlier, in my 1980 comment on Searle.

And footnote 9:

The idea that competition among many nonmental factors results in some folk-psychological conscious states may encourage another claim of Dennett’s, that the search for a neural correlate of consciousness “is probably a wild goose chase” because “the processes that elevate contents to consciousness are like the processes that elevate [evolutionary] lineage divergences into speciation events” (2009, p. 234). But the analogy with evolution is questionable; the competing factors *may well* [my

italics—DCD] result in states we can taxonomize along standard folk-psychological lines, allowing for a subsequent relatively specific process that results in those states' becoming conscious.

I think this is highly unlikely, but I'll grant it is *possible*; my aim in making the analogy to speciation events was to undermine the opposite modal assumption: that, if consciousness is an entirely physical phenomenon, there *must be* neural correlates of consciousness. Not so. Speciation is, I submit, an entirely unmysterious physical phenomenon which eventually yields striking manifestations—lions and tigers and bears. Oh my—but the time and place of its onset can only be *retrospectively* and roughly estimated; there is no Biological Correlate of Speciation. (If, for instance, you attempt to pin down the necessary and sufficient conditions for speciation by some (relatively arbitrary) proportion of accumulated change in “the” genome of a species, making cross-fertility “impossible,” you will face the prospect that any species slowly going extinct must end its days in an pitiful explosion of tiny speciation events, as shrinking and isolated gene pools briefly include a majority of non-cross-fertile genomes. You cannot “see” speciation in a moment; *as with consciousness*, you always need to ask the question: And then what happens?

## Works Cited

- Dennett, D. C. (1978). Two approaches to mental images. In *Brainstorms: Philosophical essays on mind and psychology* (pp. 174–189). Cambridge, MA: MIT Press.
- Dennett, D. C. (1980). The milk of human intentionality. *Behavioral and Brain Sciences*, 3, 428–430.
- Dennett, D. C. (1991). *Consciousness explained*. New York, NY: Little, Brown.
- Dennett, D. C. (2017). *From bacteria to Bach and back*. New York, NY: W. W. Norton.
- Debner J. A., & Jacoby, L. L. (1994). Unconscious perception: Attention, awareness, and control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(2), 304–317.
- Shields, W. E., Smith, J. D., Guttmanova, K., & Washburn, D.A. (2005). Confidence judgments by humans and rhesus monkeys. *Journal of General Psychology*, 132, 165–186.

## 6.1 IS CONSCIOUSNESS A TRICK OR A TREAT?

Jesse Prinz

Consciousness is often said to be the icing on the mental cake. Having a mind is helpful insofar as it allows for adaptive, intelligent behavior, but a mind without consciousness would have about as much appeal as a silent symphony. Whether pleasant or painful, consciousness seems to provide something special—something that logs and laptops lack. Our daily pursuits are often motivated by the promise of an experiential payoff. Without this, we think, life might not be worth living.

Throughout his monumental career, Dennett has challenged readers to consider the possibility that the apparent treat of consciousness is really a trick. The common-sense understanding of experience, enshrined in esteemed philosophical theories, is deeply mistaken. He even calls it an illusion (for example, in the title of a 2007 TED talk). It is hard to imagine a more radical conclusion. That which we think we know most intimately turns out to be a mirage. One might dismiss such a claim out of hand as a kind of philosophical extravagance were it not for the fact that Dennett defends it with mastery. In fact, he comes to these issues with such verve and industry that one comes out thinking that the common-sense view is a hopeless, silly, embarrassment. He holds it up to ridicule, he stirs up puzzles, and he batters it with counterevidence. Dennett also presents an alternative theory of consciousness that is in equal measure surprising and sensible. His 1991 book, *Consciousness Explained*, might have been more accurately called *Consciousness Explained Away*, and by the last chapter, we are left ready to cast off consciousness like an old school uniform, blushing at our untutored past. After reading Dennett, one comes to think that self-respecting naturalist should put consciousness in the bin with gods and monsters—a charming idea, but one that grownups should learn to live without.

Dennett's provocations brought me to the study of consciousness. Reading him in my student days was nothing short of thrilling. He helped teach me that philosophy can play a role in shaking deeply held assumptions. He also gave me, and many others, the courage to pursue philosophy in a way that makes contact with the cognitive sciences. Progress can be made if we abandon the presumption that mentality is special, and rather look for its place in the natural world. At the same time, I began to wonder whether such domestication requires denunciation. Dennett approaches consciousness like a fin de siècle skeptic who reveals that the local soothsayers' séances are shams. There is another kind of naturalist project that is more reductive than destructive. It takes common-sense and rehabilitates it by mapping the manifest image onto underlying mechanisms. Dennett has often played this role. He is an apologist for free will, a believer in human rationality, and defender of the view that morality can exist without supernatural underpinnings. He is also a long-time advocate of empirical research on mentality. It's hard to imagine the science of consciousness flowering as it has without his abiding influence and input.

Here my goal is to examine some of the battlegrounds on which Dennett has fought for his thesis that consciousness is an illusion. I will consider three psychological phenomena where Dennett has argued that common-sense theories of consciousness get things badly wrong. I will also consider his critique of the concept of qualia. In response to each of these cases, I will argue that common sense is not as far off-base as Dennett implies. But that does not mean common sense gets it right. In consciousness studies, we are sometimes asked to choose between views that present consciousness as a magical phenomenon and those, like Dennett's, that present it as a conjuring trick. In *Sweet Dreams*, Dennett contrasts real magic and the trickery used by magicians (Dennett, 2005, p. 58). He associated belief in real magic with dualists and also with ordinary folk who puzzle over how consciousness could rise in the brain. His own position likens consciousness to a bag of tricks. This metaphor implies that consciousness is radically unlike what it seems. I think we should abandon belief in magic, and I have little patience for dualism. I even think it's something of an embarrassment to philosophy that such views continue to be taken seriously. At the same time, I think Dennett's talk of tricks may undersell resources available to the naturalist.

The position that I will defend is not intended as a refutation of Dennett. It incorporates some of the most important lessons of his work. Everything I've ever written on consciousness carries a debt to Dennett. On a more personal note, he has supported my career since the earliest days, and he has been a mentor to my students. He works tirelessly to advertise the achievements of researchers whose efforts might have otherwise gone under the radar in professional philosophy.



Dennett has been a profoundly positive force in making the kind of philosophy that I do possible. For that, and much else, I am extremely grateful.

## 1. Pictures in the Brain

Among naturalists in the study of consciousness, there are two main orientations: demystification and debunking. Demystifiers try to save the phenomenon. They take an alleged aspect of conscious experience and try to show that our ordinary way of understanding it is importantly right. Debunkers challenge common sense. They show that we are subject to systematic confusions or delusions about experience. Dennett is widely regarded as a debunker, and this tendency was in place as the very start of his career. In *Content and Consciousness* (1969), which was first submitted as his Oxford University doctoral dissertation, Dennett takes aim at several core aspects of what he takes to be the common-sense conception of consciousness: it is not a locus of control, he argues, nor is it the locus of active, creative thought. He also challenges the accuracy of introspection and, most memorably, he denies that mental images can be understood as mental pictures. It is with this last chestnut I'd like to begin, since the debate about mental imagery beautifully illustrates the contrast between debunking and demystification.

Dennett's attack on the common-sense view of mental imagery focuses on the idea that visual images are like pictures, viewed by the mind's eye. In place of the pictorial view, Dennett advances the suggestion that mental images are more like linguistic descriptions. He points to analogies between images and descriptions and to corresponding disanalogies between picture and words. He also seeks to explain why we might be fooled into thinking that mental images are picture-like. These moves are paradigm examples of debunking. We are told that a mental phenomenon is not really as it seems and as it is ordinarily taken to be, and we are also offered an explanation for why ordinary views are so off base.

Dennett offers several examples to support his contention that mental images are more like verbal descriptions than pictures. One of these is the duck-rabbit from Wittgenstein (following Jastrow). Dennett notes that a picture of a duck-rabbit would be ambiguous, but our experience of a duck-rabbit, in perception or mental imagery, would not be. The retinal image caused by seeing a duck-rabbit is, Dennett conceded, like a picture in this respect, but within the brain it is assigned a meaning, and the resolution of the ambiguity suggests that it does not retain its picture-like format.

As a second example, Dennett notes that pictures are committal about features that can be absent in descriptions. A picture of a person will indicate whether she or he is wearing a hat, but a description need not. Mental images,



Dennett contends, are more like descriptions in this respect. We can imagine someone without specifying anything about headwear. Dennett gives another example of this kind, which he labels the Tiger and his Stripes. Pictures of tigers, he presumes, have a countable number of stripes. Not so mental images. Like a description, a mental image of a tiger may specify that it is striped without giving any indication of how many stripes are visible.

If mental images are really descriptions, why do they seem like pictures? Dennett offers two answers to this question. First, he says that mental images are like visual perception, and we confuse the thesis that imagining is like seeing for the thesis that imagining is like picturing. This confusion stems, in turn, from a confusion about the nature of seeing. We think vision provides a richly detailed representation of the world, much like a painting, but this is an illusion: vision provides descriptive information at great speed, and, whenever we think of something that might be missing from our inner representation, vision samples the external world again and provides the absent details. This gives the impression that the perceptual state had that detail to begin with. Dennett compares vision to a novel that gets written-to-order at breakneck speed. This provides a second strategy for explaining the impression that visual images are like pictures. Images may seem to be richly detailed because we can add missing bits at will.

Dennett also challenges another presupposition of the pictorial view: the idea that there is a process of inner perception. The idea that we form mental pictures usually gets paired with the metaphor of a mind's eye, which is posited to perceive them. Dennett balks at this metaphor. He says it invites a regress. Pictorialists assume that both imagery and perception use mental pictures. Therefore, if we literally perceive mental images, the mind's eye must generate a picture of the mental image, and that, too, must be pictured to be perceived, and so on. This is plainly untenable.

Armed with these strategies, Dennett is able to conclude that the pictorial theory of mental imagery is mistaken. The mental museum has captions but no pictures. Dennett also diagnoses the urge to postulate mental pictures; he thinks it rests on an elementary, but pervasive mistake. Its appeal derives from the fact that people think that mental representations must somehow resemble what they represent. A representation of a pink elephant is presumed to be elephant-shaped and pink. But there is no reason to assume this is so. Dennett advocates the following ground rule when trying to explain how the mind works:

*The Content/Vehicle Principle:* Don't assume that representational vehicles have the very properties that they represent.

This has been a guiding principle in Dennett's work. Decades after *Content and Consciousness*, we find him stating, "In general, we must distinguish features of representings from the features of the represented" (Dennett & Kinsbourne, 1992, p. 188). Applied to imagery, it cautions against assuming that representations of picture-like contents must themselves be picture-like. Given the prevalence of this assumption in both common-sense and theoretical accounts of imagery, Dennett can be described as an imagery debunker.

When *Content and Consciousness* was written, the most ardent defenders of the pictorial view would have been non-naturalists. In philosophy, the view was associated with sense data theory, a position in the philosophy of perception that posited nonphysical mental intermediaries between mind and world, that had colors and shapes (defenders included C. D. Broad and H. H. Price). We perceive these mental entities using some kind of inner sense, and then infer the existence of an external world. In the years after the book's publication, the pictorial view found a class of defenders. Roger Shepard and Jacqueline Metzler (1988) conducted pioneering psychological experiments on mental rotation, and Stephen Kosslyn (1973) followed with experiments on image scanning. These studies aimed to show that mental images had picture-like attributes, but there was little temptation to suppose that they were nonmaterial. Kosslyn (1980) adopted a computer analogy, saying mental pictures are pictures in a *functional* sense. They are mental representations that function as if they were spatial; for example, each part of the representation corresponds to a part of the represented object, and access to the representations follows principles of adjacency, such that the sequence of access mirrors proximity relations between the represented features. Since Kosslyn's account is compatible with a materialist theory of the mind, he is best described as a demystifier. His work aims to make pictorialism compatible with scientifically grounded materialism.

Kosslyn spent the better part of three decades defending a demystified form of pictorialism. By the mid-1990s, with the advent of functional neuroimaging, he declared that the imagery debate had been resolved in favor of pictorialism (Kosslyn, 1994). Kosslyn triumphantly rejects the Content/Vehicle Principle, and concludes that, in the case of visual images, vehicles resemble their contents. This can be shown, he says, by looking at the pattern of brain activity when people form mental images. Such activity has a spatial configuration that resembles what's imagined. Should we concede, then, that Kosslyn is right and that Dennett's descriptivist view is mistaken? Should we be demystifiers here rather than debunkers? And what of Dennett's arguments?

Adjudicating this debate will illustrate a recurrent theme in the rest of this chapter. In print, I have sided with the pictorialist side of the imagery debate (e.g., Prinz, 2002), and I think there are reasons to resist some of the arguments

in *Content and Consciousness*, but closer examination reveals an important kernel of truth in Dennett's critique of inner pictures, and the most defensible position may lie somewhere between debunking and demystification.

Let's start with Dennett's arguments. Some of these are difficult to assess because they seem to depend on intuitions that may not be shared. Can you really visualize a person's head without settling whether the head has a hat on it? I confess I cannot. Sometimes I wonder whether Dennett's images are just different from mine. I have art training, which might tend to promote especially vivid imagery, but Dennett is a skilled draftsman himself, so I can't imagine that this accounts for the difference. I honestly don't know what he's imagining when he says he can visualize a person's head without knowing whether there is anything on top of it. Perhaps he is conjuring up an image that is cropped, distorted, or abstracted. For example, the cubist, Juan Gris, has a 1916 portrait of his partner, Josetta, at the Reina Sophia Museum in Madrid, and one cannot tell whether she is depicted as wearing a hat. Such examples show that imagery can be noncommittal about such things, but they provide a further reason to doubt Dennett's example. If he insists that his images are noncommittal in these ways, it does not follow that they are nonpictorial. They may just be somewhat abstract.

Things are a little different with the Tiger and his Stripes. One doesn't need a cubist rendering to see that pictures of a tiger can lack determinacy. I've never seen a photo of a tiger with a countable number of stripes, since tiger stripes are broken and irregular in ways that defy counting (try to count the stripes in Figure 1). There is no settled principle of stripe individuation for tigers. Moreover, were we to count the stripes in a tiger photo, we'd easily lose



FIGURE 1. Count the stripes.

track. Counting takes effort, and double-checking is often required. In visual imagery, this problem is greatly compounded. It takes effort to maintain an image, and that may make it difficult to simultaneously count image parts, and when counting begins, the image will fade, making it hard to double check. We might find that after counting stripes in one part of the image, we would have to reconstruct other parts of the image, with no guarantee that the regenerated parts retained their original number of stripes. Thus, even if tiger images did have a fixed number of stripes, it might be impossible to count them. There is no reason to suppose, therefore, that mental images of tigers have countable stripes even on the assumption that such images are pictorial.

Duck-rabbits raise another set of issues. Dennett claims that, unlike pictures, a mental image of a duck-rabbit is not ambiguous, and he further implies that resolution of ambiguity, in perceiving duck-rabbit figures, must be achieved using nonpictorial means. Both claims can be challenged. Finke, Pinker, and Farah (1989) conducted a study in which they showed that people can reinterpret mental images. For example, they asked people to form a mental image of a capital D and a capital J, and then to mentally rotate the D, counterclockwise 90 degrees, and place it atop the J. People then discover they have imagined an umbrella. Peterson, Kihlstrom, Rose, and Glisky (1992) showed that many people can reinterpret mental images of duck-rabbit figures (they memorize the figure thinking it has just one meaning and are then asked to discover another by examining their image). Chambers and Reisberg (1992) found that some people have difficulty doing this, but they also discovered a reason for that: when memorizing the duck-rabbit people seem to attend to features most relevant to a current interpretation, leaving other features either vague or distorted in the direction of that interpretation. Thus, the mental images they form are less ambiguous than the original figure. This is not because mental images are descriptive, but because we sometimes translate ambiguous visual inputs into more determinate mental representations—which, in this case, may mean more determinate imagery.

Dennett's efforts to explain away pictorial theories can also be challenged. The fact that visual imagery is like vision would not, in and of itself, explain why imagery seems pictorial. It simply pushes that question back, as Dennett acknowledges. Kosslyn would chime in here noting that the brain's visual pathways use topographical organization well after the retina, suggesting that spatial configuration is literally, not just functionally, implemented in the brain. Against such evidence, Dennett's descriptivist theory of vision—his second strategy for explaining away pictorialism—begins to look implausible. Visual topographical maps may preserve much of the information in the retinal array. Indeed, they could contain much more, since these representations can be sustained across

multiple saccades, allowing the visual system to compile a relatively rich representation of the external inputs (just how rich will be considered in the next section). The idea that vision is writing a novel is not ruled out by vision science, but there is an attractive alternative: photosensitive retinal cells create a kind of pixel map of the world, which then get transduced into neural representations that preserve and enrich this structure.

Dennett's most powerful argument against pictorialism may be his critique of the mind's eye. Here Kosslyn would gladly concur. Pictorial images in the mind-brain are not perceived. Rather, they are processed by consumer systems that make use of their pictorial qualities. Systems that, for example, can scan across a topographic map sequentially, or enhance a subsection, much in the way we might focus our eyes on an external detail.

Given the way the imagery debate unfolded in the years after *Content and Consciousness*, we might be inclined to say that Dennett's debunking position was wrong, and the demystifiers were right. Dennett might even concur. In *Consciousness Explained* (1969, chapter 10), he concedes much to Kosslyn. He grants there may be picture-like (or, he prefers, diagram-like) representations used in mental imagery. Dennett's openness to empirical correction is refreshing. But he doesn't go all the way. He speculates that mental diagrams may be supplemented by language-like symbols to increase efficiency. Still, Dennett's position seems more demystifying than debunking.

I now want to suggest that Dennett may have gone slightly too far in his concessions to Kosslyn, or at least, he leaves the reader with that impression. There are a number of ways in which mental images cannot be like pictures. First, they lack color, since there is not actually coloring in the brain. Second, they may represent depth differently. A picture is a two-dimensional array, and as such, rotation along the  $x/y$ -axes should be computationally much easier than along the  $z$ -axis, in depth. In  $x/y$  rotation all features of the array remain constant except orientation, but  $z$ -axis rotation transforms features and reveals hidden parts. Mental rotation studies show that people are often as fast at  $z$ -axis rotation as  $x/y$ -rotation, which would be inexplicable if mental images were just pixel maps. Third, mental images may represent spatial features, such as length, differently from pictures. Lines in a picture have measurable length. Not so in mental pictures. Consider the image formed when looking at a Mueller-Lyer illusion. The lines of the actual figure can be measured. The lines of the image cannot, and, indeed, we can't even say which of the two lines is distorted. The image represents the lines as different, but we can't say whether one line is shrunken or the other is stretched. We could say this in a drawn copy of the illusion.

Kosslyn's claim that there are topographical representations in the brain offers little help here. The topography is grossly distorted. Brain regions that

correspond to the retinal array are grossly skewed, and more neural real-estate is dedicated to the fovea than to the periphery. These “images” are also upside down and duplicated (with variations) in numerous brain areas spread across two hemispheres. They are also encoded in electrical pulsations of neurons, not actual lines. Multiple neurons may represent any given point in the input, and neurons corresponding to adjacent points don’t make physical contact; they are linked by chemical signals. One might say the points are electric and the links between them are chemical. The Content/Vehicle Principle seems to be observed after all.

Somehow a lattice of pulsating neurons is able to function as a mental image. The brain does seem to exploit spatial arrangements, which is an analogy with pictures, but the analogy stops there. Pictorialism pushes too far. That doesn’t mean descriptivism is right. Brainese is no more like words than it is like pictures: there are no sounds, like spoken words, and there is little reason to think that the primitive elements have anything akin to orthography. It’s not even clear that Brainese is compositional, since it makes extensive use of population coding rather than decomposing into discrete symbols. When we rely on analogies to pictures and words, we inevitably misrepresent how representation works in the nervous system. If we were to develop a biologically plausible model of how mental imagery works, both pictorialism and descriptivism would be exposed as seriously flawed theories. What, then, should we say about demystification and debunking? Would the neutrally plausible account vindicate common sense or undermine it?

This is a confounding question for two reasons. First, it treats the contrast between demystification and debunking as dichotomous. Kosslyn may be right that some aspects of pictorialism are preserved, but others are not. So we get a position between these extremes. That, incidentally, seems to be the right reading of Dennett’s concessions in *Consciousness Explained*.

A more difficult issue concerns the relationship between Brainese and conscious experience. The fact that brains encode information in surprising ways does not entail that phenomenology is not what it seems. Perhaps all those geometric distortions are somehow resolved in experience. To infer weird phenomenology from weird wiring may be to confuse levels of analysis. Here, we quickly enter some troubled waters that Dennett has been traveling for decades. Levels, one might be tempted to think, are real strata of nature, like the sedimentary levels studied in geology. But what license do we have to reify phenomenology in this way? Why think it is anything over and above the activity measured in the brain, and what might it be if it were? This is an issue I will return to later. Dennett’s writings often imply an answer that sounds radically debunking, and his naturalist critics often sound demystifying. I will suggest, again, that the truth may lie in between.

## 2. From Richness to Rags

The next theme that I want to take up is already forecast by Dennett's treatment of imagery: the idea that consciousness is less rich than common sense makes it out to be. By the 1990s, various research programs put empirical pressure on the assumption that visual experience is rich. But, rather than concluding that visual experience is sparser than traditionally thought, many mainstream scientists try to find alternate explanations that would save its apparent richness at all costs. Dennett, of course, is unmoved by such machinations. A long-time proponent of the view that richness is illusory, he has had little patience for those who try to preserve this tenet of common sense. Here I want to give a progress report on that debate.

The primary locus of Dennett's intervention is the phenomenon of "filling in." It has been well established that the eyes are not capable of delivering a richly detailed snapshot of the external world. Sharp focus and rich color saturation are possible only in the fovea, which subtends a minuscule visual angle, and the eyes also make several saccades each second, resulting in a jittery input. To make matters worse, each eye has a blind spot located in the region of visual space where the optic nerve joins the retina (the optic disk). Oddly, most people don't seem to notice these imperfections. Most would probably describe vision as relatively stable, highly detailed, uniformly colored, and lacking any gaps. This creates a puzzle: Why does our impression of what vision is like depart so radically from the deliverances of the eyes? The standard answer in vision science is that visual systems in the brain compensate by filling in missing details. Such compensation might include sharpening and supplementing peripheral inputs and adding visual content to our two blind spots based on surrounding information.

Dennett thinks that this story rests on an elementary mistake closely related to the one diagnosed in the Vehicle/Content Principle. Those who fail to abide by the Vehicle/Content Principle assume that a representation of pink things must itself be pink. There is a corresponding assumption that says, if something is missing in a representation it must be represented as absent. Dennett repudiates this assumption. He advances another ground rule:

*The Absence Principle:* An absence of representation does not entail a representation of an absence.

Those who think that the visual system fills in missing information are guilty of violating this principle. They implicitly assume that the absence of color receptivity in peripheral vision would mean we represent the periphery as black and white, or, that the absence of receptors around the optic disk would mean that



we represent the visual world as containing two holes. If the Absence Principle is true, there is no need to compensate for these limits in optical inputs. The visual system does not represent anything as missing; it simply doesn't comment on certain things.

Dennett illustrates his position with some colorful examples. In *Consciousness Explained*, he asks readers to imagine being in a room with wallpaper covered by Andy Warhol-inspired portraits of Marilyn Monroe. Foveal vision would be able to focus on no more than one of these Marylins at a time, yet we would represent the walls as covered with Marylins. For those who flout the Absence Principle, the temptation is to say that the visual system takes its sharp foveal representation of Marilyn and then copies it throughout the visual field. Dennett thinks this is extravagant. A more efficient solution would be to simply introduce a representation that says, in effect, "there are more Marylins" (Dennett, 1991, p. 354).

In a subsequent discussion, Dennett (1996) introduces another example. Imagine seeing a realist painting of an urban landscape that represents some people walking in the distance (Dennett describes a work by Bernardo Bellotto, p. 166). On close examination, these people turn out to be little daubs of paint with no discernable features. The impression we have when viewing the whole composition is not of mere blobs, as we might experience in viewing an abstract painting, but of accurately rendered figures strutting along. How is this possible? Fans of filling-in will be tempted to say that the visual systems add details to the paint daubs. Dennett thinks this is absurd. The visual system does not represent the distinct features as featureless; it simply doesn't comment on their features. So there is no need to apply a mental paintbrush.

It is hard to deny the appeal of Dennett's in-principle argument, but it would be hasty to deny that filling-in occurs. For one thing, some experimental work strongly suggests that the visual system adds information to gaps and blindspots in the visual field. For example, Ramachandran (1992) was able to demonstrate that when a small occluder is placed on a display that is showing a twinkling pattern, the occluder is rendered invisible after a few seconds, and if the twinkling pattern is suddenly replaced by a uniform gray field, people will then experience a twinkling pattern where the occluder once was. This suggests that a neural representation of the pattern spreads across the region of visual space that is occupied by the occluder, and that this endogenously generated pattern endures a few seconds after the surrounding area has been changed. Further evidence has been obtained using cellular recording in the visual cortex of monkeys. Fiorani, Rosa, Gattass, and Rocha-Miranda (1992) found that cells in the monkeys' primary visual cortex (V1) corresponding to the region of the blind spot react to stimuli presented in the surrounding visual fields, suggesting that they interpolate based on nearby activity. Similarly, De Weerd, Gattass, Desimone, and Ungerleider,



(1995) found that cells in V2 and V3 that have artificially induced blind spots in their receptive fields respond in ways that conform to surrounding cells (for a more recent review, see De Weerd, 2006).

Dennett, of course, is aware of such results. He is not moved by them because he thinks the presence of a visual response in blind spots tells us little about the nature of that response. For example, in Ramachandran's twinkling experiment, we cannot know whether the cells themselves are twinkling—that is, whether there is a cell corresponding to each twinkle (Dennett, 1996, p. 166). The brain may use a sparser more symbolic code that indicates “more twinkling here!” without re-presenting every twinkle. As Akins and Winger (1996) point out, the presence of activity in these early visual areas does not settle the debate about filling-in, since visual awareness might reside at latter stages of process or at multiple stages or at different stages, depending on context. A rich representation in V1 might be registered by a much sparser representation later on, and awareness might be implemented by that sparse successor.

I am disinclined to follow Dennett and Akins down this route. Though I can't review it here, I think there is very powerful evidence for the view that consciousness arises relatively early in the visual stream—not at V1, but in areas such as V2 and V3 that have been called “intermediate-level vision.” Following Jackendoff (1987), I have argued elsewhere that visual experience is restricted to the intermediate level (Prinz, 2012). The evidence includes response profiles, time-course data, and lesion studies. Moreover, I think intermediate-level representations are relatively rich, which is to say there is a mapping between the cellular response and the precise details of the input. Responses are also highly sensitive to orientation, scale, hue, luminance, and other features of the input, which counts against the view that they operate at a high-level of abstraction. Cells here don't respond to twinkling as such, but to very precise kinds of twinkling. Given the evidence for activation in blind spots at this level of processing, I think it's reasonable to say the filling-in account has the upper hand.

I also think we should resist some of the examples that Dennett uses to motivate his position. The example of “more Marilyns” seems to saddle the proponent of filling-in with the view that repeating patterns are represented in full detail. But it would not be an affront to common sense to suppose that patterns are experienced with varying resolution, such that focally presented Marilyns might fade outward into blurry ones as one moves toward the periphery. The impression of richness is not an impression of sharpness. Rather it is an impression of something like repleteness: when glancing at a patterned wallpaper, one has the impression of forms throughout the entire visual field. This is entirely consistent with what we know about retinal inputs and processing in intermediate-level vision. Peripheral areas are usually active, even when we are focused on the center.

A sparse representational system might behave differently. The representation, “More Marylins” could be instantiated once, rather than repeated in every corner of the field, so the presence of meaningful activity in cells corresponding to the whole field favors the idea that representation is rich, rather than sparse.

Further evidence for richness can be obtained from studies of change blindness. In these studies, people are presented with two consecutive images and asked to report the difference; results demonstrate that even dramatic changes go unnoticed (Simons & Rensink, 2005). On the face of it, change blindness suggests that we store only a tiny fraction of the visual information that is presented to us. This would be resounding support for Dennett’s sparseness view. On closer examination, however, the study of change blindness actually supports visual richness. For example, there is evidence that the unnoticed changes still exert an influence on priming, and they can be successfully selected in forced-choice paradigms (Silverman & Mack, 2001; Mitroff, Simons, & Levin, 2004). This suggests that details are seen but forgotten.

In his own discussions of change blindness, Dennett resists the inference to sparse representations. In his most detailed discussion (to my knowledge), Dennett (2005, pp. 82–91) does not come down firmly on the side of those who say unnoticed changes are not represented. Rather, he focuses on a more specific question: Are they consciously experienced? Here, like a Pyrrhonian skeptic, he raises difficulties for every answer. His aim is to show that we lack incorrigible knowledge of conscious states. I stand firmly in Dennett’s corner on that issue. Still, I think we can avoid skeptical positions on the status of unnoticed changes. To settle this question, we can look for a well-confirmed theory of when consciousness arises and then see whether unnoticed changes meet the conditions specified by that theory. My own view—which I’ve defended in Prinz (2012)—is that attention is necessary and sufficient for consciousness. Unnoticed changes sometimes fall outside the ambit of attention; even when we focally attend to specific parts of a display, we often allocate diffuse attention to surrounding areas. Diffusely attended items are conscious, but they may not be encoded in working memory, so reportable knowledge of their presence is not sustained from one moment to the next. On other occasions, attention might be so occupied by the search for a change that it withdraws fully from some regions of the display. In these cases, the attentional theory of consciousness entails that the unnoticed changes are not experienced, though they may still be capable of priming. Thus, change blindness turns out to be not one phenomenon but two: inattentional blindness, which eliminates consciousness, and fleeting experience, which does not. I think Dennett would agree that the status of unnoticed changes could be settled by a well-confirmed theory of consciousness, and he is even sympathetic to the attentional theories, as we will see below.

For present purposes, the crucial point concerns richness. If we take the cases where unnoticed changes are diffusely attended but not encoded, we might conclude that experience is rich. The impression of richness is accurate. (Notice, I am not endorsing Ned Block's, 1995, claim that there can be consciousness without access. With Dennett, I think diffusely attended stimuli are accessible—they can impact deliberation and report, even if they happen not to.) What about cases where attention is narrowly focused, and features of a display disappear? Here I am inclined to say that experience is sparser than it would ordinarily be, but not in a way that offends common sense. Within the narrow focus of attention, details may be richly represented. The main impact is a shrinking of the visual field. That is a perfectly familiar phenomenon. While reading this, for example, the visual expanse around your computer screen or printed page may have shrunk considerably, and as soon as you stop reading, it may appear to expand.

Such considerations may seem like bad news for Dennett, given his reoccurring claim that visual experience is less rich than ordinarily presumed. But I think the verdict isn't quite so clear. Dennett may be wrong to say that richness is an illusion, but core aspects of Dennett's debunking arguments remain untouched.

Consider, first, Dennett's Absence Principle. Even if visual repetition is rich, this principle may stand: when something is missing from the visual input, it does not entail that we see it as absent. Researchers who draw such an inference are making a mistake. A recent example can illustrate. We all blink about 15 times a minute, but we don't notice these blinks. Why? One answer is that the visual system fills in the time interval of the blink. Working with this hypothesis, Gawne and Martin (2002) set out to find that stage of processing at which vision ignores blinks. They compared blinks to externally caused darkenings, which we do not ignore. One might expect that some stage of visual processing would treat these two conditions differently: shutting down during external darkenings, but maintaining activity during blinks. Gawne and Martin did not find such a difference, and they demonstrated a dampening in activity during blinks throughout the visual hierarchy. This supports a Dennettian moral: we should not assume that a dampening in the brain will engender a dampening in experience.

Dennett is also right to assume that filling in has its limits. When we see letters on a page or a screen, as in a page of text, we assume that we are experiencing each letter as a letter. This does not seem to be the case. For example, de Gardelle, Sackur, and Kouider (2009) presented arrays containing letters and pseudo-letters and found that participants believed only letters were present. Similar findings have been obtained for decades using "eye-movement contingent displays." In this method, the text on a page is surreptitiously replaced when it is not being read (Rayner & McConkie, 1976). When ordinary words are replaced by pseudo-words or x's, the change typically goes unnoticed. The window of

awareness is about 14–15 letters to the right of the currently fixated word, and 3–4 to the left. Beyond that, fairly dramatic changes are rarely detected. Readers in these studies will insist they saw a page of text even when staring at a page that is mostly x's. Might they be filling in these x's for real words? If so, which words? It seems unlikely. This is not exactly the same as the sparse representation view, which would say we represent the periphery as “more words” without specifying how they are arranged, but it does suggest that we are content to treat non-words as more words, and that is tantamount to a kind of illusion. We think we are seeing more detail than we are.

The verdict for richness is not as dire as Dennett would sometimes have us believe. There is evidence that the visual world is replete with information, arranged throughout the visual field, and that gaps may be filled in. Vision is not a fragmented or tattered rag, but an opulent, richly embroidered expanse. Still, we may systematically overestimate the specificity of these riches. What seems like more letters (or more Marilyns) might be something quite different (imagine a wall of Marilyns that tossed in a few Nixons and Maos). There is a gap between the contents of experience and what we take ourselves to be experiencing. This doesn't quite debunk richness, but it destabilizes the common-sense conviction that we know what we are seeing. Overall, the picture fits nicely with Dennett's career-long campaign against first-person authority.

### 3. Shouting Contests

I turn now to a final case study in this tour of Dennett's illusionism. It involves one of the parade of examples from *Consciousness Explained*: the color phi phenomenon. In color phi, a display flickers between two slightly offset colored disks; if parameters are just right, it will look like one object moving back and forth and changing hue as it moves. Dennett notes that there are two stories about how this mind-brain generates this moving object. One possibility (dubbed Orwellian) is that we experience the two differently colored disks as independent entities, but then form a memory as of a single moving object, forgetting how it has seemed milliseconds earlier. Another possibility (dubbed Stalinesque) is that there is delay before the onset of consciousness, so we never experience the disks as independent, but only experienced the fused moving object that is generated by the mind a moment after seeing the display. Dennett claims that we cannot decide between these options. He uses this case to diagnose what is wrong with prevailing theories of consciousness and to advance his own positive theory.

Dennett thinks that the temptation to think that we must choose between the Orwellian and Stalinesque interpretations of the color phi phenomenon

is based on a mistake. In particular, it is based on the assumption that there is some kind of “Cartesian Theater,” where conscious experiences come together in the mind-brain. On most theories of consciousness, there is either a physical location in the brain where consciousness happens, or else a set of functional conditions that are necessary and sufficient for consciousness. If so, then for any mental state, we could settle whether or not it is conscious. We could ask if the two disks in the color phi display meet the conditions for consciousness before a representation of their fusion is generated. Dennett rejects the underlying assumption. Using another metaphor, he adopts the following position:

*No Turnstiles Principle:* There is no single place, physically or functionally, that marks the boundary between conscious and unconscious states.

If this principle is right, then it could turn out that there are many things going on in the mind-brain, in different places at different stages of processing, that have equal claim to being called conscious. Each can be thought of as a stream of consciousness, or at least a font of consciousness, and each presents a different version of what is happening now. There is no single theater, because there is no audience: no self to whom one might deliver the privileged results of “subpersonal” processing. Instead, each mind is a vast bureaucracy of different agencies working somewhat independently, with no main office.

This picture of the mind also provides the basis of Dennett’s positive view of consciousness. In *Consciousness Explained*, he called it the “multiple drafts theory,” but he now focuses on another metaphor: cerebral celebrity, or fame in the brain (cf. Dennett, 2005). The basic idea is that the many mental agencies are each vying for control of certain systems, such as the systems that underlie self-ascription or verbal reporting. This can be thought of as a kind of shouting contest. Each office of the mind tries to take over the public-address (PA) system to broadcast their latest achievements. On any given occasion, it may turn out that one of these offices has commandeered the PA system more vociferously than others, and it gets heard more clearly than all the others. That is fame in the brain. And the office that enjoys this fleeting good fortune can, in that moment, exert an influence over self-ascription and verbal report. “Did you see one disk or two?” asks the experimenter, and, then, depending on what else happens to be going on, the office of visual motion detection might gain control and issue its report, “one; it was colorfully shuffling back and forth.”

This is a deflationary theory of consciousness because it doesn’t posit anything like phenomenality: a special, ineffable, fundamentally subjective quality

of experience. Many different offices can have their chance at fame in the brain, and gaining fame is just a matter of being heard. No office is privileged and no magical transformation occurs when the shouting contest is won.

The No Turnstiles Principle was already in place when Dennett wrote *Content and Consciousness*. It is more a reflection of his philosophical convictions than a consequence of some empirical results. Suppose that one had taken the issue of turnstiles as a research question, rather than as an assumption one way or the other. One might have asked, let's take a class of states that would be widely regarded as conscious across a wide range of theoretical orientations, and see if they have anything functionally or physically in common. For example, one could begin with states that we are able to freely report as being experienced.

My own view is that if we begin this way, we actually will arrive at a kind of turnstile view (Prinz, 2012). As mentioned earlier, I think evidence supports the contention that consciousness arises when and only when we attend. When attention is withdrawn from a stimulus, people report not having experienced it, and when attention is applied, it becomes reportable. When attention is diffusely distributed, we can accurately report that a stimulus was present, but we can't necessarily name it. Attention, in turn, can be given a coherent functional characterization. There are different theories of attention, but my favorite says that attention arises when sensory information becomes available for encoding in working memory. I also argue that the mechanisms of attention operate at a specific stage of sensory processing, the aforementioned intermediate level. These mechanisms also have a monolith physiological description involving specific kinds of neural oscillations. The intermediate level, structures that control attention, and working memory are distributed in the brain: there is no consciousness center. But they are functionally cohesive and thereby qualify as a kind of virtual turnstile.

The account just described, which I call the AIR theory for Attended Intermediate-Level Representations, does not directly deliver a verdict on color phi, but it provides some resources for doing so. Empirically, we have reason to believe that motion in color phi is not a false memory but an actual representation generated by the brain. Neuroimaging studies have established that in the phi phenomenon, activity is generated in the region between the two disks, confirming that the brain is filling in the path between them to create a motion trajectory (Larsen, Madsen, Lund, & Bundesen, 2006). Behaviorally, it has been shown that the motion generated by phi displays can mask an item placed between the two disks (Yantis & Nakama, 1998). Chong, Hong, and Shim (2014) went even further, showing that such masking is especially powerful when the item between the two disks has a color that is intermediary

between them. This evidence strongly suggests that the mind-brain generates a representation of the movement in color phi, including the transition in color. There is also good reason to think this endogenously generated representation of movement is conscious. After all, people report seeing the movement and the empirical results correspond closely to what people report seeing. This supports the Stalinesque view. A defender of the Orwellian view might insist that the two disks are consciously experienced as separate entities just prior to the experience of their unity as a moving object, and this prior experience is simply forgotten. This seems unlikely, however, since people do not report having such an experience, and consciousness ordinarily results in working memory access, which in turn, is ordinarily adequate for reportability.

In recent writings, Dennett has independently developed ideas that align with core aspects of the AIR theory. He has been exploring the thesis that attention is crucial for consciousness (Cohen & Dennett, 2011), and he shares my view that attention allows access to working memory. He has even expressed sympathy for the possibility that consciousness is restricted to the intermediate level (Dennett, 2015). This suggests that Dennett is moving closer to a turnstiles view, and, thus, further from a deflationary view.

This transition looks like a major break from the debunking spirit of Dennett's earlier work, but I want to end this section with a reason for thinking that a crucial part of the earlier view is preserved, even if something like the AIR theory is right. Consciousness has a functional turnstile on the AIR theory insofar as we can specify function and physiological changes that take place when and only when a state is conscious. There is always a fact of the matter regarding which states are conscious, and being conscious is not just a matter of being reportable. This goes against claims in *Consciousness Explained*. There is, however, a wrinkle that hasn't been revealed. The mind-brain may be a bit like an urban metro system with different turnstiles for each sensory modality. Granting that attention confers consciousness, there is no guarantee that intermediate-level representation in different senses (or even different sensory features in the same sense) will reach the attentional threshold concurrently. Sensory systems process information at different speeds and operate using proprietary coordinate systems. These can sometimes be aligned through coordinate transformations, but that may not happen all the time (Pouget, Ducom, Torri, & Bavelier, 2002). There is evidence, too, that we can have different streams of attention (e.g., Johnson & Zatore, 2006; Keitel, Maess, Schröger, & Müller, 2013), and working memory is not a unified center that represents information to anything approximating a Cartesian ego or self. So Dennett's vision of a bureaucratic mind that can have multiple conscious agencies and fluctuate in degree of integration is wholly consistent with AIR theory.



#### 4. Trick or Treat?

Throughout his career, Dennett has likened consciousness to a trick. As with stage magic, it appears to be marvelous, mysterious, and incompatible with physical laws. This is an illusion. We can account for consciousness without positing anything supernatural. With this, all naturalists would agree. Efforts to accommodate consciousness within scientific approaches to the mind presuppose that consciousness is not as mysterious as it may seem. Within the naturalist camp there are, however, divisions. Some try to preserve much of our common-sense understanding; such authors (the demystifiers) presume that many features of consciousness are as we take them to be. They are content to postulate picture-like images and rich contents, for example, and they tend to suppose that consciousness arises because of some functionally or physically cohesive process. Others (the debunkers) like to challenge common-sense assumptions, and they tend to offer deflationary theories. Dennett has been the preeminent figure in this camp. He thinks common-sense views about consciousness, and philosophical theories that enshrine common sense, should be dispelled. Dennett has also resisted the idea that there is a functionally or physically coherent process that leads to consciousness in the mind-brain. The only thing conscious states share on Dennett's cerebral celebrity view is fleeting control over an inner PA system. In this sense, consciousness is little more than a story we tell ourselves.

In looking at three examples from Dennett's oeuvre, I've tried to defend a position between demystification and debunking. I've been a demystifier since my earliest days in philosophy, but Dennett has helped me see that consciousness must actually be quite different from how it seems. Does this mean that consciousness is just a trick? Is the AIR theory really deflationary in the way the cerebral celebrity is presumed to be? To address these questions, I want to meditate on a term that has been conspicuously absent in the foregoing discussion: qualia. Those who think consciousness is a treat rather than a trick cling to the idea that conscious states have qualitative character. Is there anything to that claim?

This is a tough question, burdened by the fact that "qualia" is a somewhat obscure technical term. Dennett (1992) says that qualia are, by definition, "ineffable, intrinsic, private, directly apprehensible ways things seem" (p. 47). One might quibble with this definition, but I think it's a pretty good starting place. By way of conclusion, I want to briefly suggest that AIRs (attended intermediate-level representations) may have each of these features, in some sense. If so, maybe we don't need to be so skeptical about qualia.

Let's begin with ineffability. If consciousness arises at the intermediate level in sensory processing systems, then conscious states will often be encoded at a level of abstraction that is appreciably finer than words. The specific shades and



shapes we see have no names. They could be described with great labor, but such tasks are rarely undertaken, and the descriptions we would arrive at would differ from the shapes and colors we experience in important ways: they would have different syntactic components, and understanding these would require mastery of concepts that are not needed to have a visual experience. In these respects, intermediate-level representations have no linguistic synonyms or translations: they are ineffable.

Intermediate-level representations are also, in some sense, intrinsic. This point requires the introduction of some details of the AIR theory that I have not yet mentioned. Suppose we ask what distinguishes a red experience from a blue experience? The answer cannot be that they are caused by red and blue things respectively since the causes are external and may already be absent by the time the experience arises—think of a briefly flashed display. Nor can these color experiences depend on the outputs they cause. Imagine training yourself to call red things “blue,” (and conversely); this would not cause a dramatic reversal in the way these colors seem. From a neurocomputational perspective, the best account of how we distinguish red and blue experiences is that there is a difference in their neural implementations. Elsewhere I argue that each color is implemented by a different neural vectorwave (Prinz, 2012; see also McClurkin, Zarbock, & Optican, 1996). If so, the difference between red and blue is intrinsic. This is not to say we can be conscious of them without accessing them. A vectorwave in a petri dish would not be conscious. But the means by which we access these two states is the same: both come into consciousness with attention. What sets them apart is their intrinsic properties.

Privacy is a slightly more complicated matter. Dennett’s (1992) characterizes this alleged attribute by saying that “interpersonal comparisons of these ways of appearing are (apparently) systematically impossible” (p. 46). So stated, this condition may be too difficult to meet for anything that isn’t supernatural, since states and processes in the natural world can generally be measured, and hence compared. The vectorwaves of two individuals are potentially comparable, for example. But I think Dennett’s gloss on privacy may be too stringent, even for a dualist qualiophile. After all, contemporary dualists believe that qualia have neural correlates in the actual world, so one might be able to determine that two individuals have similar experiences by examining their brains. Moreover, it’s easy to imagine a science-fiction thought experiment in which two brains get wired together allowing two people to see what things are like from another point of view.

I think the privacy condition is better captured by thinking about the comparison with things that are public. Two points of contrast deserve emphasis here. First, public things are external to the mind, so that when mental states share

public content, they can be characterized as representing the same feature in the external world. Second, public things can be fully known via different means; epistemic access does not depend on a specific way of knowing. These features can be called externality and objectivity. In contrast, private things are internal and subjective. On this characterization, AIRs can be regarded as private. Unlike some “intentionalist” theories of phenomenal character, AIRs are not individuated by external referents. In this respect, they differ too from what Dennett (1992) calls “phenomenal information properties”—his favored alternative to qualia (p. 71). AIRs are individuated by dynamic neural patterns. One can measure and describe these patterns from a third-person perspective, but there is also a sense in which they are subjective. Knowing the neurocomputational description is no substitute for being in one of these states. The neurocomputational description tells us about these states using public language, but such descriptive knowledge is neither necessary nor sufficient for knowing what vectorwaves are like to experience. It is not necessary, because we all experience our vectorwaves prior to knowing anything about neuroscience, and it is not sufficient, because having a vectorwave instantiated is tantamount to having an experience, and reading about a vectorwave is not. Experiencing is a way of knowing it on just about anyone’s account, so there is something left out by the verbal description. Vectorwaves are, in this sense, subjective.

This last point bears on the direct knowledge condition: qualia are supposed to be known directly. Many forms of knowledge are indirect: the thing known and the means by which we know it come apart. We learn about the external world by having internal representations, for example. Not so with vectorwaves. On the theory under consideration, instantiated vectorwaves are experiences. If experience is a species of knowing, then we can be said to know our vectorwaves by having them. The gulf between the thing known and the way of knowing collapses. This is a kind of direct knowledge. Once again, I don’t mean to suggest that conscious states can occur without being accessed. Vectorwaves are conscious when and only when they are modulated by the brain processes underlying attention and those modulations make them available to working memory. Consciousness requires access. But this does not mean we know vectorwaves indirectly. We don’t know them by representing them. As Dennett likes to point out, there is no reason for the brain to re-present a state that is already taking place within the brain. Access is not a matter of representation; it is a matter of informational efficacy. Access gives vectorwaves a chance to impact downstream processes.

For these reasons, I think the AIR theory is compatible with realism about qualia. AIRs satisfy the job description that Dennett lays out in his analysis of how the term “qualia” is used. Given Dennett’s recent sympathies with both attention and the intermediate level, perhaps he should soften his stance on

qualia. Dennett's move toward something like the AIR view may also help vindicate qualia-talk in another way. There is a further reason for resisting the naturalization of qualia implicit in much of Dennett's work. Dennett has long been a pluralist about consciousness. Dennett (2005) sometimes compares consciousness to the "tuned deck," a card trick that can be performed in several different ways so no one can ever guess how it's done (p. 72). This pluralism spells trouble for qualia. Without any underlying unity, it makes little sense to think that there is a class of special mental states, spanning across the senses, that are in some sense alike (alike in having qualitative character) and different from all others. This source of resistance is removed, however, if we can show that there is, in fact, a single process unifying qualitative mental states. Dennett indicates some continuing allegiance to pluralism in his recent adoption of the attention view; he says there are multiple kinds of attention, "distributed, featural, spatial, internal and so on" (Cohen & Dennett, 2011, p. 360). But this pluralism may turn out to be superficial. Attention can be controlled in different ways and allocated to different things, but the underlying processes may be the same (Prinz, 2012). If so, a neat mapping between qualitative character and a kind of brain process would be possible. Perhaps, then, qualia can be demystified rather than debunked.

In keeping with the theme of this chapter, however, I want to end by suggesting that the demystification of qualia can go only so far. AIRs may satisfy the qualia job description, but that doesn't mean there is nothing tricky about them. In fact, qualia may be the brain's greatest trick. I said that qualia are, in some sense, known directly. We know them by having them. On the other hand, this direct knowledge fails to inform us of their actual nature. Qualia do not reveal themselves to be vectorwaves. One cannot, through introspection of qualia, infer that conscious states are brain states. Furthermore, the similarity relations that we find in our qualia are not derived from their intrinsic qualities, even though we think they are. Red and blue experiences are vectorwaves and so are circular experiences, experiences of sounds, smells, and textures. These things all strike us as very different, but their neural correlates are relatively similar, and the physiological ordering that exists between them may have little to do with the similarities we report. We also make systematic errors about their structure. For example, red seems more basic (or "primary") than orange, but the vectorwave is equally complex. Such misjudgments presumably derive from the fact that beliefs about similarity and structure are not simply read off qualia, but rather depend on much unconscious information as well. Once we give up on the idea that the intrinsic nature of qualia can be known by a "reading off" process, we are left with a view that allows for radical error about the nature of our own experiences.

The moral about qualia echoes what I said about mental images, filling in, and color phi. Dennett has suggested that widely entrenched beliefs about these things are wildly mistaken. He has engaged in decades of debunking. He has even suggested that consciousness itself is an illusion. In each case, I suggested that Dennett might go too far. Naturalism is compatible with pictorial images, rich visual representations, and a fact of the matter about how we experience the color phi. As we have just seen, naturalist theories can vindicate qualia as well. But when we look more closely, we find that the naturalist accounts on offer do not leave our entrenched beliefs entirely untouched. In some way, the scientific image is a radical departure from what we believed about ourselves prescientifically. Thus, we arrive at a mixed verdict. Consciousness is not a trick; it is not an illusion. Those who believe in qualia and describe consciousness as the icing on the mental cake can find some comfort in scientific results. But consciousness is a tricky treat. We treasure our conscious experiences, but we are often profoundly mistaken about them. For me, this is one of the most important and striking lessons of Dennett's work.

## Acknowledgments

I am indebted to Felipe De Brigard and Bryce Huebner for extremely helpful comments and corrections. Their work gives me great hope for the future of philosophy. I am also grateful to Daniel Dennett, who has been a *sine qua non* for me in many ways.

## Works Cited

- Akins, K., & Winger, S. (1996) Ships in the night: Churchland and Ramachandran on Dennett's theory of consciousness. In K. Akins (Ed.), *Perception* (pp. 173–197). Oxford, UK: Oxford University Press.
- Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(02), 227–247.
- Chambers, D., & Reisberg, D. (1992). What an image depicts depends on what an image means. *Cognitive Psychology*, 24, 145–174.
- Chong, E., Hong, S. W., & Shim, W. M. (2014). Color updating on the apparent motion path. *Journal of Vision*, 14(8), 1–12.
- Cohen, M., & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, 15, 358–364.
- de Gardelle, V., Sackur, J., & Kouider, S. (2009). Perceptual illusions in brief visual presentations. *Consciousness and Cognition*, 18, 569–577.

- De Weerd, P. (2006). Perceptual filling-in: More than the eye can see. *Progress in Brain Research*, 154, 227–245.
- De Weerd, P., Gattass, R., Desimone, R., & Ungerleider, L. G. (1995). Responses of cells in monkey visual cortex during perceptual filling-in of an artificial scotoma. *Nature*, 377, 731–734.
- Dennett, D. C. (1969). *Content and consciousness*. New York, NY: Routledge.
- Dennett, D. C. (1991). *Consciousness explained*. Boston, MA: Little, Brown.
- Dennett, D. C. (1992). Quining qualia. In Marcel, A. J., & Bisiach, E. E. (Eds.), *Consciousness in modern science* (pp. 42–77). New York, NY: Oxford University Press.
- Dennett, D. C. (1996). Seeing is believing—or is it? In K. Akins (Ed.), *Perception* (pp. 158–172). Oxford, UK: Oxford University Press.
- Dennett, D. C. (2005). *Sweet dreams: Philosophical obstacles to a science of consciousness*. Cambridge, MA: MIT Press.
- Dennett, D. C. (2015). The friar's fringe of consciousness. In I. Toivonen, P. Csúri, & E. van der Zee (Eds.), *Structures in the mind: Essays on language, music, and cognition in honor of Ray Jackendoff* (pp. 371–378). Cambridge, MA: MIT Press.
- Dennett, D. C., & Kinsbourne, M. (1992). Time and the observer: The where and when of consciousness in the brain. *Behavioral and Brain Sciences*, 15, 183–201.
- Finke, R. A., Pinker, S., & Farah, M. J. (1989). Reinterpreting visual patterns in visual imagery. *Cognitive Science*, 13, 51–78.
- Fiorani, M., Rosa, M. G. P., Gattass, R., & Rocha-Miranda, C. E. (1992). Dynamic surrounds of receptive fields in primate striate cortex: A physiological basis for perceptual completion. *Proceedings of the National Academy of Sciences (USA)*, 89, 8547–8551.
- Gawne, T. J., & Martin, J. M. (2002). Responses of primate visual cortical neurons to stimuli presented by flash, saccade, blink, and external darkening. *Journal of Neurophysiology*, 88, 2178–2186.
- Jackendoff, R. (1987). *Consciousness and the computational mind*. Cambridge, MA: MIT Press.
- Johnson, J. A., & Zatorre, R. J. (2005). Attention to simultaneous unrelated auditory and visual events: Behavioral and neural correlates. *Cerebral Cortex*, 15, 1609–1620.
- Keitel, C., Maess, B., Schröger, E., & Müller, M. M. (2013). Early visual and auditory processing rely on modality-specific attentional resources. *Neuroimage*, 70, 240–249.
- Kosslyn, S. M. (1973). Scanning visual images: Some structural implications. *Attention, Perception, & Psychophysics*, 14(1), 90–94.
- Kosslyn, S. M. (1980). *Image and mind*. Cambridge, MA: Harvard University Press.
- Kosslyn, S. M. (1994). *Image and brain: The resolution of the imagery debate*. Cambridge, MA: MIT Press.
- Larsen, A., Madsen, K. H., Lund, T. E., & Bundesen, C. (2006). Images of illusory motion in primary visual cortex. *Journal of Cognitive Neuroscience*, 18, 1174–1180.

- McClurkin, J. W., Zarbock, J. A., & Optican, L. M. (1996). Primate striate and prestriate cortical neurons during discrimination: II. Separable temporal codes for color and pattern. *Journal of Neurophysiology*, 75, 496–507.
- Mitroff, S. R., Simons, D. J., & Levin, D. T. (2004). Nothing compares two views: Change blindness can occur despite preserved access to the changed information. *Perception and Psychophysics*, 66, 1268–1281.
- Peterson, M. A., Kihlstrom, J. F., Rose, P. M., & Glisky, M. L. (1992). Mental images can be ambiguous: Reconstrual and reference frame reversals. *Memory and Cognition*, 20, 107–123.
- Pouget, A., Ducom, J. C., Torri, J., & Bavelier, D. (2002). Multisensory spatial representations in eye-centered coordinates for reaching. *Cognition*, 83(1), B1–B11.
- Prinz, J. J. (2012). *The conscious brain*. New York, NY: Oxford University Press.
- Ramachandran, V. S. (1992). Filling in gaps in perception: Part 1. Current Directions in *Psychological Science*, 1, 199–205.
- Rayner, K., & McConkie, G. W. (1976). What guides a reader's eye movements? *Vision Research*, 16, 829–837.
- Shepard, S., & Metzler, D. (1988). Mental rotation: Effects of dimensionality of objects and type of task. *Journal of Experimental Psychology: Human Perception and Performance*, 14(1), 3–11.
- Silverman, M., & Mack, A. (2001). Priming from change blindness [Abstract]. *Journal of Vision*, 1, 13a.
- Simons, D. J., & Rensink, R. A. (2005). Change blindness: Past, present, and future. *Trends in Cognitive Sciences*, 9(1), 16–20.
- Yantis, S., & Nakama, T. (1998). Visual interactions in the path of apparent motion. *Nature Neuroscience*, 1(6), 508–512.

## 6.2 REFLECTIONS ON JESSE PRINZ

Daniel C. Dennett

Trick or treat. Exactly. Jesse Prinz's title captures the guiding thread of my various campaigns against inflated—we might well call them *mystified*—views of consciousness. There is nothing mysterious about why we might *think* that consciousness was the most wonderful, miraculous, important phenomenon in our lives; without consciousness, after all, our lives wouldn't matter at all. Nicholas Humphrey (2006, 2011) argues that this belief is itself an evolved adaptation, a prospect that should be grounds enough for a vigorously skeptical approach. Before we can hope to make progress on the scientific study of consciousness we have to get a grip, and recognize that we are *not* authorities on the properties of the (neural) events that constitute (or “subserve” or just “are”) our conscious experiences, even if we have a trivial sort of authority (the authority of authorship of our reports) over the *contents* of our experiences. This follows from what Prinz calls my Content/Vehicle Principle:

Don't assume that representational vehicles have the very properties that they represent. (p. 174)

I applaud Prinz's articulation and elevation of this as my “guiding principle,” though I don't think I have ever made just that claim. I wish I had, because it simplifies the construction of the vise I use to squeeze the recalcitrant mystifier. It is entirely uncontroversial that there are many kinds of representations that do not have the properties they represent—“purple ball gown” isn't purple, “ $x^2 + y^2 = 16$ ” isn't round, and “sharp aroma of cinnamon” doesn't smell like cinnamon (sniff the page if you doubt me)—so you should never just *assume* that a representation has the properties it represents. How, then, do you know your conscious experiences have the properties you are eager to attribute to them?

Please close your eyes and imagine, as vividly as you can, a bright red capital letter T, in a sans-serif font, on a black background. Done? Now imagine the same T in bright blue. Done? Something changed in your brain, in virtue of which you stopped representing a red T and began representing a blue T. What? Nobody knows—yet. But we can be sure this is *not* a case of a red T-shaped thing turning blue! I have a pair of powerpoint slides on a flash drive that represent these two Ts, and we *do* know what physical properties of those two data structures differ: some memory cells storing the first slide have an electron where the corresponding memory cells storing the second slide do not. (The presence or absence of an electron in a memory location is the flash drive's way of representing a 0 or 1, and the codes for blue and red are different bit strings.)

Now it might have been that our brains stored visual information much the way cinema film frames do, with red regions representing red regions and so forth, but as an empirical matter we can be quite sure that this is not the brain's way. It's dark in there, and without a source of light to "extract" the information from the representation when needed, a red region in the brain is as inert a representer as a bit string would be in a cinema projector. That is beside the point, you may say, since you are not talking about the properties of things like external surfaces, but the properties of internal, *mental* states, *phenomenal* properties. Phenomenal red needs no light to shine and phenomenal trumpet blast cannot be picked up by any microphone inserted in the brain. Phenomenal properties are, apparently, whatever they have to be to have the effects on belief and action that they do, but that leaves it wide open that phenomenal redness might *be* a bit string in a data structure, for instance. The bitstring in the flash drive is a representation of redness in virtue of the systematic way it generates activities that ultimately produce red spots in some external medium (inkjet print, LED pixel on a screen, etc.) The neural whatever that represents red does *not* generate a display somewhere. It represents red in virtue of the systematic way it is produced by the reception of light reflected off red objects and the systematic way it modulates belief and behavior about the colors of things seen. No Double Transduction! (See Andy Clark's chapter (7.1) and my Reflections (7.2) for more on this.) If you say that such informational or computational properties could not *be* the phenomenal properties you are talking about, I want to know what makes you think there are any such phenomenal properties at all?

When I suggested, way back in *Content and Consciousness* (Dennett 1969), that vision was not a series of pictures and more like a novel that gets written to order at breakneck speed, that was a rather desperate fallback, but in those days, we didn't have any well-developed alternatives. John Maynard Keynes was once asked, "Do you think in words or pictures?" His reply: "I think in thoughts." Good move! But not very informative. As Prinz says, "When we rely on analogies



to pictures and words, we inevitably misrepresent how representation works in the nervous system.” All these years I have eagerly awaited the theoretical enlightenment that would allow me to give a more positive answer, and this is beginning to emerge. Prinz speaks cautiously of “a lattice of pulsating neurons” that is able to function as a mental image by exploiting spatial arrangements, and this is homing in on the same mother lode of new models of neural representation that has attracted the attention of the Churchlands, Thomas Metzinger, and other empirically minded philosophers.

These are promising developments precisely because they glide right past the traditional stopping point (phenomenal properties, or qualia) and go on to show how they can deal with what I call the Hard Question: And then what happens? For instance, the believers in “phenomenal consciousness” insist on its “richness” in contrast with mere “access consciousness,” and they think they can somehow directly discover this richness in their experience. In “What Is the Bandwidth of Perceptual Experience?” Cohen, Dennett, and Kanwisher (2016) argue that *visual ensembles* and *summary statistics* provide a novel format of neural representation that can account for the (false) impressions we have of richness while accounting for what we *can do* (and then what happens?) with the information provided in this format. Our article thus responds directly to Prinz’s plausible but mistaken claim, “Vision is not a fragmented or tattered rag, but an opulent, richly embroidered expanse” (p. 185). “The key idea here is that the visual system exploits the redundancy found in real-world scenes to represent a large amount of information, often extending into the visual periphery, as a single summary statistic.” (Cohen et al., 2016, p. 325).

Prinz’s account of the “more Marilyns” phenomenon is, basically, the ensemble statistics account:

The example of “more Marilyns” seems to saddle the proponent of filling in with the view that repeating patterns are represented in full detail. But it would not be an affront to common sense to suppose that patterns are experienced with varying resolution, such that focally presented Marilyns might fade outward into blurry ones as one moves toward the periphery. The impression of richness is not an impression of sharpness. Rather it is an impression of something like repleteness: when glancing at a patterned wallpaper, one has the impression of forms throughout the entire visual field. (p. 182)

The Cohen et al. (2016) paper illustrates the idea of ensemble statistics with a *metaphorical* representation of what we see, rendered as a wildly distorted picture that nevertheless honors the ensemble statistics. We are not claiming that what

we see are such distorted images but that what we see cannot distinguish one of these distorted images from another. The representation itself is neither words nor pictures and hence can't be rendered in a journal article. The figures in the article were composed by an algorithm that generates one or another of the kazillions of different images that have same ensemble statistics. That a representation (in a "lattice of pulsating neurons" for instance or a "vectorwave") of this ensemble statistic could yield the same answers to "and then what happens?" questions is demonstrated by the fact that this randomly selected alternative image with the relevant statistical properties is indistinguishable under the relevant perceptual conditions from the undistorted image

So Prinz and I are on the same page, but there are a few continuing disagreements worth mentioning. Prinz is still dubious of my forays into experiments in imagining. "Can you really visualize a person's head without settling whether the head has a hat on it? I confess I cannot." I agree with Prinz that if one sets out to visualize a person's head without going into the hat issue, one will fail, or end up with some lame obscuring element, but that is the wrong sort of case to try. Try these instead:

Imagine an angry cluster of about eight adults, some men, some women, approaching you on an otherwise deserted street, a few brandishing sticks, a few holding rocks they seem about to throw.

Imagine this in as much detail as you can muster—perhaps you're a novelist writing up a key episode, or even a screenplay writer going to the trouble of fixing many of the telling details of the shot. Done? Now, ask yourself did any or all of the women and men have hats on? You probably *can* answer that question, since you were primed to go into that detail, so let's turn to a few other questions: Were any of the women wearing skirts/dresses? heavy boots? Did those who were brandishing sticks hold them in their right or left hand? Were any of the rock-wielders left-handed? Any beards among the men? Anybody wearing glasses?

Why couldn't imagination be a kind of "neglect in normals." In hemi-neglect, patients just don't notice what's missing on the left side of space. The pictures they draw are bizarre, and it's hard to believe that they can't *see* how bizarre the pictures are. But much experimental testing confirms that they are not malingering or insincerely denying that they see any problems. I suggest that imagining should be seen not so much as *abstract*, as Prinz suggests, as *negligent*: we seldom go into all the detail and don't even keep track of the details we neglect. That could mean that folks sometimes negligently "draw" *pictures in their heads*, but if you can be that negligent, you could also fail to notice that the products of your efforts weren't pictures at all, but something nameless, however familiar.

One further strike against the pictures in the head idea should appeal to Prinz, my fellow draftsman: I am sure we both know the frustrating, even baffling, feeling of setting out to draw something not from nature—a still life or figure study or portrait—but from imagination alone—and having in mind “*exactly*” the elegant shape we intend to sketch, but being embarrassingly unable to render *that* on the drawing paper. (I would love to be able to sketch, freehand, the sinuous bodies and flowers of Mucha, for instance, or Dali’s impromptu sketches—which in my mind dwarf everything else Dali did—but while I can “picture” the vibrant soaring lines in great detail, I cannot execute the intentions represented there. But put a Mucha print in front of me to copy and I can do fairly good job, if I take my time and concentrate.)

There seems to me to be an unnoticed contradiction between Prinz’s judgments about two of my favorite phenomena, change blindness and color phi. About change blindness he writes:

For example, there is evidence that the unnoticed changes still exert influence of priming, and can be successfully selected in forced-choice paradigms (Silverman & Mack, 2001; Mitroff, Simons, & Levin, 2004). This suggests that details are seen but forgotten. (p. 183)

Or: it suggests that these are cases of unconscious priming. There is no doubt that the unnoticed changes get represented at several lower perceptual levels; does that count as “seen but forgotten”? And about color phi he writes:

A defender of the Orwellian view might insist that the two disks are consciously experienced as separate entities just prior to the experience of their unity as a moving object, and this prior experience is simply forgotten. This seems unlikely, however, since people do not report having such an experience, and consciousness ordinarily results in working memory access, which in turn, is ordinarily adequate for reportability. (p. 188).

What about change blindness favors the Orwellian view, which is so firmly dismissed in the case of color phi? Either I am missing something or Prinz is actually providing support for my claim that there is no principled way of distinguishing Orwellian and Stalinesque accounts of some (not all) phenomena.

Finally, one caveat about “filling in.” As with my words-versus-pictures dichotomy in 1969, my “filling-in”-versus-“no-filling-in” discussions have tended to suggest an all-or-nothing view, but in fact I am happy to acknowledge many cases of partial, or *sorta* filling in (In 1991 I did discuss a few; see p. 349). Prinz cites Ramachandran’s twinkling and artificial scotoma phenomena, which are very

interesting but in fact demonstrate that such filling in *cannot* occur in normal vision. Unless you deliberately fixate your view on a target, immobilizing your eyes, they dart about in saccades, with pauses of no more than 300 msec. It takes thousands of milliseconds for these phenomena to develop, guaranteeing that they are not responsible for any sense of a plenum you have in normal vision.

Prinz's review of my arguments about these cases is a valuable update, yielding the conclusion that, even if, as he shows, I oversimplified or overstated the case for my initial provocative claims, there remains something unsettling that I am right about. And he adds empirical support that is new to me, such as the Gawne and Martin (2002) blinking experiments, which excellently illustrate my Absence Principle.

## Works Cited

- Cohen, M., Dennett, D. C., & Kanwisher, N. (2016). What is the bandwidth of perceptual experience? *Trends in Cognitive Sciences*, 20, 324–335.
- Dennett, D. C. (1969). *Content and consciousness*. New York, NY: Routledge.
- Gawne, T. J., & Martin, J. M. (2002). Responses of primate visual cortical neurons to stimuli presented by flash, saccade, blink, and external darkening. *Journal of Neurophysiology*, 88, 2178–2186.
- Humphrey, N. (2006). *Seeing red: A study in consciousness*. Cambridge, MA: Harvard University Press.
- Humphrey, N. (2011). *Soul dust: The magic of consciousness*. Princeton, NJ: Princeton University Press.
- Mitroff, S. R., Simons, D. J., & Levin, D. T. (2004). Nothing compares two views: Change blindness can occur despite preserved access to the changed information. *Perception and Psychophysics*, 66, 1268–1281.
- Silverman, M., & Mack, A. (2001). Priming from change blindness [Abstract]. *Journal of Vision*, 1, 13a.

# 7.1

## STRANGE INVERSIONS

### PREDICTION AND THE EXPLANATION OF CONSCIOUS EXPERIENCE

Andy Clark

One of Daniel Dennett's favorite things is the "strange inversion." Strange inversions occur when things work in ways that turn received wisdom upside down. Hume offered us a strangely inverted story about causation, and Darwin, about apparent design. Such accounts often display what seem to be puzzling causes (intelligent design, experienced causal connection) as less-puzzling effects—the effects, respectively, of blind selection and of experienced succession. An especially important species of strange inversion occurs, Dennett suggests, when we project our own reactive complexes outward, painting our world with a litany of elusive properties such as cuteness, sweetness, blueness, sexiness, funniness and more. Such properties strike us as experiential causes, but they are (Dennett argues) really effects—a kind of shorthand for whole sets of reactive dispositions rooted in the nuts and bolts of human information processing. Understanding the nature and origins of that strange inversion, Dennett believes, is thus key to understanding the nature and origins of human experience itself. I examine this claim, paying special attention to recent formulations (Dennett, 2013, 2015) that link that strange inversion to the emerging vision of the brain as a Bayesian estimator, constantly seeking to predict the unfolding sensory barrage.

#### 1. The Trouble with Double Transduction

Daniel Dennett has waged a decades-long war against obfuscatory appeals to "qualia" in discussions of the nature and possibility of conscious experience. At the heart of the many and varied skirmishes that

this has involved, however, lies a single basic imperative. It is the imperative to avoid positing “double transduction” in the brain. It is this, Dennett claims, that provides the breeding grounds for the more noxious understandings of “qualia”—intrinsic properties of experience that seem to resist satisfying forms of scientific and philosophical explanation.

So what is “double transduction”? “Double transduction” involves positing an appealing, but (Dennett claims) unnecessary, step in the flow of influence linking worldly states to human responses. Impinging energies, such as light hitting the retina, carry information about the external world (including our own bodies) and those energies impact the neural economy so as to yield spike trains and other neural states. So far, so good. Double transduction would occur if those, in turn, needed to be translated or transduced into some other form (“qualia”) for inspection and consideration before yielding verbal responses and/or gross motor actions. It is this image, Dennett claims, that is fatally flawed. There is no need for the second transduction, since the first transduction poised incoming information to do whatever needs to be done, by yielding neural states (spike trains) that are already apt to guide responses of all kinds. And of course, there is then no need then for a third transduction (back into spike trains again) either. Instead, it’s spike trains (etc.) *all the way*. As Dennett (2015) recently put it:

It is still extremely tempting to imagine that vision is like television, and that [neural] spike trains get transduced “back into subjective color and sound” and so forth, but we know better, don’t we? We don’t have to strike up the little band in the brain to play the music we hear in our minds, and we don’t have to waft molecules through the cortex to be the grounds for our savoring the aroma of bacon or strawberries. There is no second transduction. And if there were, there would have to be a third transduction, back into spike trains, to account for our ability to judge and act on the basis of our subjective experiences. . . . But biology has been thrifty in us: it’s all done in the medium of spike trains in neurons (section 1)

Dennett’s alternative picture is austere. We are hit by a barrage of energies that sense organs transduce into spike trains, etc. These are all the brain gets, and they must be “processed” so as to yield more spike trains (etc.) that drive apt actions—including verbal ones. There is simply no role available for transduction into something else (those naughty qualia, played on the inner movie screen) along the way. No second transduction. No need for an inner viewing. Instead, the brain takes the impinging energies, manifesting as rank upon rank of inter-animated spike trains, and uses them (suitably “elaborated, processed, reverberated, reentered,

combined, compared, and contrasted” (Dennett, 2015, section 2) to drive action, including verbal report.

The threat of double transduction is nicely illustrated by Dennett’s CADBLIND thought experiment (see Dennett, 1991, pp. 290–293). CAD (computer aided design) systems are able to help human designers answer questions such as “what will this object look like if rotated 70 degrees to the left” and “will this part of the object be visible when viewed from such-and-such a location?” CAD systems can do this based on coordinate maps and transformation algorithms. Such a system might generate a 3D image of an object as it would appear from some perspective *Y*. The results of such transformations are displayed on a screen for the human user, who can then inspect the screen and issue the answer (perhaps “yes, the object-part is indeed visible when the viewer is located thus-and-so”).

Now imagine a CAD system designed to be used by blind engineers. This is CADBLIND (Mark I). As a first try, the designers might retrofit the old system with a new back-end device (Dennett dubs it the “Vorsetzer”) that inspects the screen after the rotation algorithm has been applied and uses artificial intelligence (AI) vision techniques to determine, from that new raw data, whether the object-part is indeed now in shot. But clearly, this is suboptimal. For all the information needed to make that call is already in the system (indeed, it was required to form rotated image in the first place). The correct solution is thus to cut out the Vorsetzer and deliver the verdict directly on the basis of the previously processed information.

Dennett thinks that the philosophers’ explanatory appeal to inner qualia is like the posit of an inner Vorsetzer—an unnecessary add-on that misleads us into thinking that there remains (even after all the information processing chores are accomplished) some special kind of work that still needs to be done.

We can take this further. Dennett goes on to imagine CADBLIND Mark II. This second-generation machine applies the same logic to the case of color. The device first codes objects positioned in front of a camera according to their (appropriately complex) local surface reflectance properties. Thus “post-box red” (to use a rather British example) might be coded in memory as Reflectance #179 (Red 179). Now suppose we place a red hat in front of the camera. CADBLIND Mark II codes it as Reflectance #173 (Red 173). With just a little more programming, the device is now in a position to deal with questions such as “which red is deeper?” Once again, there is simply no need to generate a colored display for some inner Vorsetzer to inspect. Instead, it is set up to subtract 173 from 179, getting 6, and then to answer, “the hat is a little deeper red.”

The moral is that the path from input to judgment need not have qualia as an intervening variable. The cycle is not [Sense-Process-Qualia-Judge] but

simply [Sense-Process-Judge]. In that latter sequence, there is neither need nor room for intervening mysterious qualia (no room for mental “pigment” in which the inner colors are actually rendered—see Dennett, 1991, chapters 11 and 12).

Continuing the story, Dennett suggests we might ask a more fully language-enabled version of the machine how it knows which is the deeper red. At that point it might just say “I don’t know how I know, it just *looks* deeper. I compared the two reds in my mind’s eye and got the answer.” Qualia, we are invited to conclude, certainly *seem* to play a role in the generation of our own answers to such questions. But (just as in the case of CADBLIND) perhaps that seeming is all there is.

These are important arguments, but they seem to fall just short of delivering the goods. For isn’t there still room to ask, in the case of CADBLIND Mark II, whether it *really does seem* to it as if it is comparing shades of red in its mind’s eye? Dennett sometimes dismisses this as an almost unintelligible distinction between “real” and “merely seeming” seemings. But at other times (as we’ll shortly see) he, too, seems to agree that something more is required. Furthermore, Dennett repeatedly saddles the friends of qualia with the posit of mental “figment.” But a reasonable fan might insist that even though there need be nothing red in the head when we make our judgments concerning redness, still some complexes of (non-red) neural events manage to constitute *actual experiences of color*, while others, even though they may lead to some apparent judgments concerning color, do not.

## 2. Strange Inversions

Stamping out the threat of double transduction is thus a great start. But as Dennett (2015, section 4) rightly notes, it can hardly be a satisfying finish. For things don’t, on the face of it, seem like that at all. If it’s all just spike trains doing what they do, why does it *seem to us* as if we are rotating mental images, appreciating colors and textures, or even just recognizing people, animals, and objects? If we could answer this question (“why do things seem the way they do”) to everyone’s satisfaction, that would surely spell the end of the qualia wars.

To set the scene for such an effort, Dennett first brings onstage another of his long-term loves: the Strange Inversion. This notion is best introduced by example. My personal favorite (though not, as far as I know, one used by Dennett himself) concerns bars famous for their excellent Guinness. It turns out that the major determinant of this is how long the barrel has been open. A bar’s reputation for serving good Guinness may thus be what causes the Guinness to be good, since a bar with that reputation will serve lots of Guinness, thus keeping it fresher.



We thought the reputation was due to the Guinness. But in fact, the quality of the Guinness is an effect of the reputation.

Or (to take Dennett's core example) consider Hume on causation. The standard (noninverted) story has it that we see X causing Y, and thus form the idea of causal connections. The inverted story says no, we never see X causing Y! Instead, we see Y following X, regularly, apparently inexorably. We then come to expect this regular succession, and this strong psychological expectation is projected back onto the world as some apparently perceptible "causal connection." What appears to be a property of the world itself here turns out (if Hume is right) to be a property of the observer. In this way "we misinterpret an inner 'feeling', an anticipation, as an external property" (Dennett, 2015, section 6). In Hume's terms, we "gild and stain" natural objects with features and properties "borrowed from internal sentiment" (to quote Dennett, 2015, quoting Kail, 2007, quoting Hume, 1751/1998, Appendix 1.21).

Hume's strange inversion provides a close model for Dennett's own attempt to re-orient our thinking about qualia and qualitative experience. Thus consider, to follow one of Dennett's prime examples, the sweet taste of honey. According to the standard (noninverted) story, we like the honey because it tastes so sweet. The view from after the strange inversion is, of course, rather different. Our liking of the honey, it is suggested, is nothing but the subtle complex of reactive dispositions that honey evokes—dispositions (in many folk, at least) to seek, lick, assert to be tasty, and the like. If the inversion is on the mark, the complex of reactive dispositions comes first, and we dub things that evoke that complex of responses as "sweet." So it's not the sweetness that explains our response, it's our response that explains (constitutes) the sweetness. But not knowing this, we project the sweetness onto the world, resulting in the self-diagnosis of a perceptual realm that presents a variety of mysterious and puzzling qualia—a realm of sweet-tasting honey, cute-looking babies, funny jokes, and even (as we'll later see) red-seeming apples.

This is a good move. It aims to show why we think experience presents a world populated with all that odd stuff (sweet tastes and cute babies) when actually, what we are tracking are just more features of the world—in this case, ones invisibly intertwined with our own reactive dispositions. But exactly how does this "projection" trick work? And why does it yield seeming-qualia?

This is where Dennett adds a further (Bayesian) layer to the argument. The resulting argument structure (as schematized by Dennett, 2015, section 1) looks like this:

1. There is no double transduction in the brain.
2. Therefore, there is no second medium, the medium of consciousness or, as I like to call this imaginary phenomenon, the *ME*diuM.

3. Therefore, qualia, conceived of as states of this imaginary medium, do not exist.
4. But it seems to us that they do.
5. It seems that qualia are the source or cause of our judgments about phenomenal properties . . . but this is backwards. If they existed, they would have to be the *effects* of those judgments.
6. The seeming alluded to in (4) is to be explained in terms of Bayesian expectations.

Dennett's project is thus to account for our judgments concerning the existence of "qualia-seemings" while rejecting the idea that those judgments are effects caused by qualia-laden experiences. Instead (here is the strange inversion), the qualia-seemings are to be revealed as effects, not causes: effects rooted in our "Bayesian expectations." This is ambitious. It is the cognitive equivalent of the famous 720 Triple Flip in skateboarding, where the skater flips multiple times in vertical and horizontal space and can seem to have magically overcome gravity itself. But of course, it's not magic but physics after all.

Dennett's 720 Qualia Flip (as I'll now dub this particular strange inversion) must conjure qualia-seemings (real qualia-seemings) from what seems to be nowhere—nowhere, at least, that is already populated with qualia. The 720 flip appears to defy gravity, allowing the skateboarder to apparently gain leverage from (literally) thin air. Dennett needs to perform a similar trick with qualia. He must somehow deliver the apparent effects of qualia (the judgments we make, the things we say) without positing qualia as their cause or justification. To see how this works, we must next introduce the emerging Bayesian accounts mentioned in point 6 above.

### 3. The Bayesian Dimension

Dennett's 720 Qualia Flip is inspired by an emerging body of fundamental work on the neural roots of perception, cognition, and action. That work depicts the brain as, in essence, an organ whose job is to predict the incoming sensory stream. This image of the brain as an engine of prediction can be found in various forms in contemporary neuroscience (for useful surveys, see Bubic, von Cramon, & Schubotz, 2010; Kveraga, Ghuman, & Bar, 2007; and, for my own favorite incarnation, Friston, 2009). The full story is large and complex (for introductions, see Clark, 2013, 2016; Hohwy, 2013) but one core feature of these "predictive processing" (PP) accounts stands out as especially relevant to our current concerns. That feature is the use of acquired knowledge (in the form of a structured "generative model") to predict the current sensory input. This results in a kind of Bayesian inversion of

the standard (passive, feedforward) image of sensory processing. Instead of trying to build a model of what's out there on the basis of a panoply of low-level sensory cues, these models aim, in effect, to predict the current suite of low-level sensory cues from their best models of what's most likely to be out there (for this formulation, see Hohwy, 2007). The brain, if this story is correct, meets the incoming raw multimodal sensory stream with a complex web of learned “top-down” predictions. Its task is to generate a multilevel match for the incoming sensory data using knowledge concerning patterns and probabilities in the world. Failures to predict the shape of the sensory signal result in prediction errors that recruit new top-down hypotheses. In Bayesian terms, the task of the brain is to find the linked set of hypotheses that, given the current raw sensory data, make that data most probable. The best hypotheses will be the simplest ones that maximize the posterior probability of the observed sensory data. Notice that “observed sensory data” here just means “the impinging raw signal”—not even a hint of double transduction here!

By way of illustration, consider the simple but striking demonstration (used by the neuroscientist Richard Gregory to make this very point; see, e.g., Gregory, 1980) known as the hollow-face illusion. This is a well-known illusion in which an ordinary face mask viewed from the back (which is concave, to fit your face) appears strikingly convex when viewed from a modest distance. That is, it looks (from the back) to be shaped like a real face with the nose sticking outward, rather than having a concave nose cavity. Here, for example, is a mask of Albert Einstein, viewed from both the front (painted, convex) side and the red (unpainted, concave) side, with appropriate backlighting.

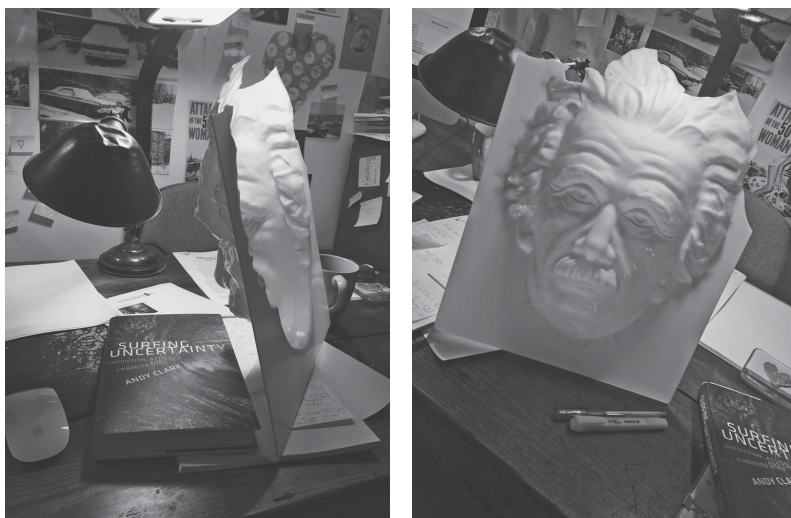
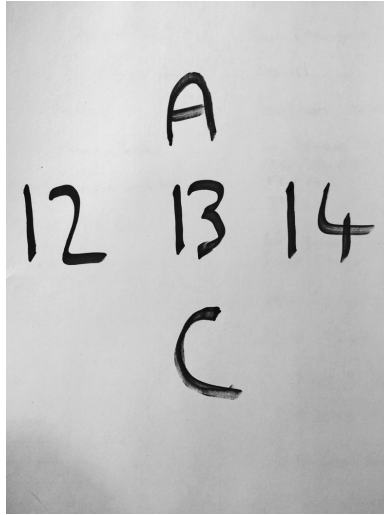


FIGURE 1. Hollow Face Illusion



**FIGURE 2.** Inverting Perceptual Processing. In the context of reading the A, the B hypothesis makes the raw visual data most probable. In the context of reading the 12, the 13 hypothesis makes the very same raw visual data most probable.

The hollow-face illusion illustrates the power of “top-down” (essentially, knowledge-driven) influences on perception. Our statistically salient experience with endless hordes of convex faces in daily life installs a deep, subpersonal “expectation” of convexness: an expectation that here trumps the many other visual cues that ought to be telling us that what we are seeing is a concave mask.

As a second example, take a look at Figure 2.

Reading left to right, the predictive brain meets the raw sensory stimulations from the central inscription with a strong expectation of the numeral 13. Reading top to bottom, the raw sensory stimulations are met with a different prediction—the letter B. In Bayesian terms, in the context of reading the 12, the 13 hypothesis makes the raw visual data most probable, while in the context of reading the A, the B hypothesis makes the raw visual data most probable (for more on this example, see Lupyan & Clark, 2015).

You might reasonably suspect that such effects, though striking, are really just some kind of psychological oddity. And to be sure, our expectations concerning the convexity of faces seem especially strong and potent, and the visual forms in Figure 1 are patently ambiguous. But if these stories are on track, the same core strategy pervades human perception. The claim is that brains like ours are constantly trying to use what they already know so as to predict the current sensory signal. Thus, take a familiar scenario such as visually perceiving the clutter of papers and coffee cups on your office desk. Glancing at your desk, a few

rapidly processed visual cues set off a chain of visual processing in which incoming sensory signals are met by a stream of downward predictions concerning the most probable states of this little world. These predictions reflect the buzzing, proactive nature of much of your ongoing neuronal processing. That torrent of downward-flowing prediction is in the business of preemptively specifying the probable states of various neuronal groups along the appropriate visual (and other) pathways. The down-flowing torrent of prediction concerns all aspects of the unfolding encounter, and is not limited to simple visual features such as shape and color. It may include a wealth of multimodal associations, and a complex mix of motoric and affective predictions. At the higher levels, it will involve knowledge and expectations concerning whole objects (such as cups) and their functions. As you visually inspect your desktop there occurs a rapid exchange (an energetic dance between multiple top-down and bottom-up signals) in which any incorrect downward-flowing guesses yield “prediction error signals.” These error signals propagate upward, and are used to leverage better and better guesses. As this process unfolds, top-down processing is trying to generate (at multiple spatial and temporal scales) the incoming sensory signal for itself. When downward-flowing (top-down) guessing adequately accounts for the incoming signal, the visual scene is perceived.<sup>1</sup>

#### 4. Completing the 720 Qualia Flip

The Bayesian/Predictive Processing story does a nice job of explaining why we resolve ambiguities in the sensory signal in specific ways, according to our (mostly

---

1. How is the knowledge used to drive the predictions acquired in the first place? Some may be innate, inherent in the basic shape of the neural economy. For example, the division of visual processing into dorsal and ventral streams, specializing in encoding object location and object identity, may reflect a kind of sedimented “knowledge” that the same objects can exist in different places. See Friston (2013). But a major attraction of the multilayer predictive processing approach is that it lends itself very naturally to powerful forms of unsupervised learning in which the attempt to predict powers the learning that makes better predictions possible. Thus, suppose you want to predict the next word in a sentence. You would be helped by a knowledge of grammar. But one way to learn a surprising amount of grammar, as work on large-corpus machine learning clearly demonstrates, is to try repeatedly to predict the next word in a sentence, adjusting your future responses in the light of past patterns. You can thus use the prediction task to bootstrap your way to the world knowledge that you can later use to perform apt prediction. Importantly, multilevel prediction machinery then delivers a multiscale grip on the worldly sources of structure in the sensory signal. In such architectures, higher levels learn to specialize in predicting events and states of affairs that are—in an intuitive sense—built up from the kinds of features and properties (such as lines, shapes, and edges) targeted by lower levels. But all that lower-level response is now modulated, moment-by-moment, by top-down predictions. For more on all this, see Hohwy (2013); and Clark (2013, 2016).

subpersonal) expectations about what is likely to be out there and how it is likely to be presented to our senses.<sup>2</sup> It also does a nice job (see Clark, 2013, 2016) of accounting for why we encounter a structured external realm, in which objects and events seem rich in meaning and significance. For the complex predictions that best explain the sensory barrage automatically present that barrage as best explained by consistent webs of hypotheses that:

1. Unearth a structured mass of interacting distal causes operating at multiple scales of space and time: cats, colors, textures, causal relations—all are on a par, inferred causes or “latent variables” inferred to account for the raw sensory flux.
2. Project the inferred causes forwards in time, so we know what the rustle of the curtain signifies for us (perhaps it heralds the appearance of our missing cat).

This is already quite a lot. We have here made progress with the question of how we come to know a structured world of distal objects, apt to display certain kinds of evolving behavior. The objects, along with their dispositions to change and evolve, are inferred as the best explanations of the waves of sensory data. This explains (in one sense) how a structured world comes into view in the first place—a world built of dogs, cats, schooners, icebergs, even mental states, such as the beliefs and intentions of other agents. All these emerge on a par, as content-items (“latent variables”<sup>3</sup>) in our best predictive model of the sensory barrage.

This, however, is the very same way that a structured world might be said to “come into view” for (to take one example from a huge and ever-expanding pool<sup>4</sup>) a contemporary deep-learning network able to recognize handwritten numerals. Thus Hinton (2007) shows how a microworld of type-identical numerals that may be inscribed in different ways may be brought into view by a species of prediction-driven learning. But such networks, though remarkable, are not plausibly ascribed “qualia-seemings.” To understand how such networks process information is not yet to reveal the origins of qualia-seemings. To put the matter starkly, using the kind of talk favored by the fans of irreducible qualia, why does it seem *like anything at all* to us when we encounter a structured realm

---

2. For example, we “expect” daytime visual scenes to be illuminated from above—an expectation that, in ecologically unusual settings, can lead to a variety of visual illusions. For some excellent discussion, see Adams, Kerrigan, and Graf (2010).

3. Latent variables are simply variables that are inferred to account for the observed (directly measured) data—which in this case is the raw sensory barrage.

4. For a useful review, see Bengio (2009).

in thought and action? This is where Dennett's 720 Qualia Flip is meant to save the day.

At the heart of the flip lies a deceptively simple observation. We are, Dennett notes, prime objects for our own prediction machinery. To grapple with our own futures, we need to predict what we ourselves will do next, what we will think next, etc. And among the things we predict about ourselves, we will find a special subclass: predictions about our own reactive dispositions. This, Dennett argues, explains why there seem to be qualia.

Consider (following Dennett, 2015) the apparent cuteness of babies. Is it our experience of the baby's cuteness that causes us to approach and want to care for the baby? Dennett's claim is that this gets things backward. It is not our experience of elusive cuteness-qualia that moves us to approach and nurture. Rather, it is our tendencies to approach and nurture (etc., etc.) that appear to us as mysterious properties of the baby, or perhaps of our experience, itself. More specifically, the claim is that the "cuteness seemings" are nothing more than learned "Bayesian" expectations concerning our own probable reactions to imminent baby-exposure. Thus:

when we expect to see the baby in the crib, we also expect to "find it cute"—that is, we expect to expect to feel the urge to cuddle it and so forth (Dennett, 2015, section 4).

The crucial move is the next one. For when we do then cuddle the baby, with all those (mostly subpersonal) expectations in place:

the absence of prediction error signals is interpreted as confirmation that the thing in the world we are interacting with has the properties we expected it to have. *Cuteness as a property passes the Bayesian test for being an objective structural part of the world we live in* [my emphasis]. (section 4)

According to this picture we (subpersonally) expect to feel like cooing and nurturing, and when those "Bayesian expectations" are met, we deem the baby itself cute. But (being the advanced, reflective agents we are) we may then diagnose ourselves as having detected something that is deeply puzzling, given our larger view of the world. For qualitative cuteness now seems to have been "given" in the input stream, just like ordinary content properties such as baby-ness (or numeral 6-ness). If this is correct, then all those apparent cuteness-qualia (and all other qualia, including redness, painfulness, etc.) are really just disguised predictions of our own web of reactive dispositions. But they are predictions that, when satisfied,



can appear to add mysterious qualitative dimensions to our daily encounters with a structured external world.<sup>5</sup>

The story on offer is meant to apply to all forms of qualitative experience, including apparently simple (“brute”) qualities such as the experienced redness of a uniform, well-illuminated, plain red surface. This is initially puzzling. Properties such as “experienced cuteness” do seem, intuitively, to involve something akin to Dennett’s “Bayesian expectations.” It seems plausible, that is, to think that at least part of what is involved in experiencing the baby as cute involves a web of reactive dispositions to approach, coo, cuddle, nibble, etc. But (at first sight) the experience of brute qualities such as redness seems different. As Dennett (2015) says:

There is no way one expects to behave in the presence of navy blue, or pale yellow, or lime green (section 7).

This appearance, Dennett argues, is deceptive. In fact, our experience of color is complex and structured, and itself involves a host of expectations concerning our own likely responses. In particular, Dennett notes that we possess many expectations concerning our own emotional responses. To demonstrate this, we are invited to imagine an agent (“Mr. Clapgras”) whose emotional responses to colors are suddenly shifted 180 degrees while leaving his cognitive (color identification) capacities intact and normal. Such an agent would be surprised to find the green grass to be anxiety provoking and unrestful, preferring the gentle calming effect of vibrant reds and shocking pinks. The very fact that such an agent would be initially surprised is testimony (Dennett argues) to the operation of a mass of mostly subpersonal predictions concerning our own likely responses to surfaces of varying colors.

At this point, Dennett’s 720 Qualia Flip is complete. The full spectrum of human qualitative experience, if Dennett is right, is best understood as a disguised appreciation of our own reactive dispositions, which (in Bayesian fashion) we project onto the world. This occurs because the incoming sensory barrage is met with predictions that track both the evolving signal and our own likely

---

5. The fans of qualia will rightly ask about that feeling of puzzlement itself. Isn’t that just another qualitative experience in need of explanation? In the end, I think whatever account we give of qualia in general needs to be deployed to explain why we sometimes judge ourselves to be feeling puzzled or surprised. Such judgments might reflect, for example, the prior low probability of a state of the world that (given some sensory evidence) is now judged the best overall hypothesis. Do not ask: but why does that feel surprising? Rather, the Dennettian suggestion must be that the tendency to make those judgements is what the “surprisingness” really consists in. (Thanks to Felipe De Brigard for some probing questions on this—see also de Brigard, 2012.)



patterns of emotional and physical response, and the lack of prediction error then invites us to treat these as specifying objective (but somewhat mysterious) aspects of the world we encounter. In reality, however, these are features not simply of the world, but of our relation to that world.

Thus unmasked, qualia do not figure in a causal chain that explains our responses and judgments. Instead (this is the strange inversion) they are explained by our own unarticulated expectations concerning those very patterns of response.

## 5. A Puzzle: “Qualia Surprise”

Dennett’s Qualia Flip gives pride of place to our expectations concerning our own reactions. For it is only by predicting our own reactions, then confirming them on exposure, that we fool ourselves (so the argument goes) into thinking that qualitative properties are being detected “out there in the world.” This version of the appeal to Bayesian expectations leads, however, to a puzzle. Before I see the baby, I may indeed expect to feel the urge to cuddle it (I may even expect to expect to feel the urge to cuddle it). But what about cases where the cuteness of something (or the emotional states associated with exposure to a new, previously never seen, color) comes as a big surprise to me? In such cases, aspects of qualitative experience seem to force themselves upon us *even though we do not predict (even unconsciously) our own reactive dispositions in advance*. Such cases (let’s call them cases of “qualia surprise”) may seem to pose a challenge to Dennett’s Bayesian reconstruction.

Consider my own recent experience of seeing (for the first time ever) a picture of an albino squirrel. If someone had asked me, before this exposure, if I expected to find albino squirrels cute, I’d have said it was unlikely. First, I don’t much like squirrels anyway. Second, other albino animals often look strange and (to my eyes) not especially attractive. I invite the reader, however, to search for “albino squirrels” online. My own experience, when first presented with one of those images, was roughly “wow, those critters *are* cute.” What are we to make of this? Did I perhaps nonconsciously predict I’d find them cute? There’s no reason, as far as I can tell, to say that. Nonetheless, exposed to that novel sensory barrage, my brain’s best (Bayesian!) hypothesis turned out to be that they were white, fluffy and—goddammit—cute.

I do not think this undermines Dennett’s core argument. But it does suggest an important nuance. The reason we can experience qualia surprise is because the raw sensory inputs here actually possess plenty of hidden structure—so spotting a novel kind of cuteness is really no harder than spotting a novel kind of chair as a chair or some crazy, futuristic automobile as an automobile. This suggests that it is our systematically structured grip on the world, realized courtesy of the neural prediction engine, that does the real work here. Importantly, that systematic grip

extends to our own physiological (interoceptively detected) responses, too, as has recently been argued by Seth (2013) and Pezzulo (2014).

The picture that I am urging thus looks like this. The brain's core task (in true Bayesian fashion) is to find the multilevel, top-down hypothesis that best predicts both the exteroceptive and interoceptive sensory flow. That hypothesis is, however, built using a highly structured palette (as realized by the hierarchical form of the generative model). The winning multilevel hypothesis then reveals both *what* I am seeing [White Fluffy Squirrel-Forms With Red Eyes—Linguistic Classification, Albino Squirrels] and, courtesy of the interoceptive elements, *how* I am responding to that seeing [Urge to Stroke, to Approach—Linguistic Classification “Cute Animals”].

This preserves the strange inversion—the perceived cuteness depends on the urge to stroke, protect, photograph, etc., rather than the other way around. But it does not depend upon my already expecting (in any nontrivial sense) to find albino squirrels cute. Instead, their cuteness is part of a hypothesis constructed on the spot so as to best accommodate the sensory (including interoceptive) data. In the future, of course, I may indeed expect to find albino squirrels cute—today's posterior is tomorrow's prior, after all! But (I am suggesting) the cuteness-seemingness doesn't have to wait for that kind of self-prediction to take root. This also preserves the *other* key aspect of the full 720 flip. For qualitative features such as cuteness here emerge as posits inferred to reduce the dimensionality of the data (including organism-related interoceptive data) in the raw sensory stream. Otherwise put, such features are just more latent variables in the generative model that best accounts for the overall (interoceptive and exteroceptive) sensory barrage, appearing to us as distal causes on a par with dogs, cats, hats, and hurricanes.

What role then remains for the second-order predictions highlighted in Dennett (2013, 2015)—for example, our expectation that we will find the baby cute, hence (on this analysis) our expectation that we will expect to want to cuddle it, etc.? Such second-order predictions, by being borne out in practice, may indeed serve to cement our conviction that we are here dealing with real features of the world, by further obscuring the importance of our own reactive dispositions. Moreover, such second-order predictions enable us to model ourselves and other agents using simple schemas that ascribe qualitative experiences to ourselves and to others. At that point, the (partly self-fulfilling<sup>6</sup>) explanatory successes of our folk-psychological practice may add further fuel to the qualiaphile's fire. By turning up in our best model of ourselves and of other agents, qualia seem earn their place in the pantheon of explanatory posits that enable us to meet the sensory stream with apt predictions.

---

6. See Zawidzki (2008) on mindshaping.

## 6. Conclusions: Realism Without Double Transduction?

The most striking thing about Dennett's 720 Qualia Flip is that superficially disparate elements (such as doghood, catness, loudness, redness, and cuteness) are all treated in much the same way. Brought together under the accommodating umbrella of structured Bayesian predictive models, these all emerge as content items (latent variables) inferred as part of the generative model that best predicts the raw sensory flow. Crucially, that sensory flow now includes a flow of predicted interoceptive information and response, and it is those aspects of the flow (reflecting our own reactive dispositions) that are responsible, if Dennett is right, for much of the apparent puzzlingness of qualitative experience. Such a story, though still sketchy and incomplete, strikes me as highly promising. It links the puzzlingness of qualia directly to a plausible story about how, and why, they seem to exist. To complete the story, it would be necessary to include the role of attention, and to say something about the relations between all this and agent-level choice and action.<sup>7</sup> But those are tasks for another day.

I want to end, however, with a question. If redness and cuteness pick out real patterns in the combined (interoceptive and exteroceptive) sensory barrage, just like dogness, what remains of the claim that qualia are in some sense unreal: that they are mere qualia-seemings? For on the Bayesian model we discover (infer) redness and cuteness and merlot-tasty-ness just as we discover (infer) dogness, hatness, and hurricaneness. To be sure, dogness here depends rather less on our human reactive dispositions (to approach, to stroke) than cuteness. But that seems like a small detail, given that each is now constructed in the same way, and charged to play the same kind of role—that of meeting the incoming sensory barrage with an apt flow of top-down prediction. Reconstructed through the lens of Bayesian prediction, qualia-seemings reflect patterns that are as real as any others. Might this be an unexpected victory for a modest (no double transduction) form of qualia realism? One way or another, the time seems ripe to start to unravel the metaphysics of the Bayesian brain.

## Acknowledgments

Special thanks to Daniel Dennett, Susan Dennett, Patricia and Paul Churchland, Dave Chalmers, Nick Humphrey, Keith Frankish, Jesse Prinz, Derk Pereboom, Dmitry Volkov, and all the students from Moscow State University who, in

---

7. I make a small start on some of this in Clark (2016). See also Prinz (2012); Hohwy (2012).

June 2014, discussed these (and many other) themes on an unforgettable boat trip among the icebergs of Greenland. Thanks to Felipe De Brigard and Bryce Huebner for stimulating comments on an earlier draft. This work was also partly supported by the AHRC-funded Extended Knowledge project, based at the Eidyn research center, University of Edinburgh.

## Works Cited

- Adams, W. J., Kerrigan, I. S., & Graf, E. W. (2010). Efficient visual recalibration from either visual or haptic feed-back: The importance of being wrong. *Journal of Neuroscience*, 30, 14745–14749. <https://doi.org/10.1523/JNEUROSCI.2749-10.2010>
- Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2, 1–127.
- Bubic A., von Cramon, D. Y., & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, 4(25), 1–15.
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36, 181–204.
- Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. New York, NY: Oxford University Press.
- De Brigard, F. (2012). Predictive memory and the surprising gap: Commentary on Andy Clark’s “Whatever next? Predictive brains, situated agents, and the future of cognitive science.” *Frontiers in Psychology*, 3, 420.
- Dennett, D. C. (1991). *Consciousness explained*. Boston, MA: Little, Brown.
- Dennett, D. C. (2013). Expecting ourselves to expect: The Bayesian brain as a projector. *Behavioral and Brain Sciences*, 36, 209–210.
- Dennett, D. C. (2015). Why and how does consciousness seem the way it seems? *OpenMIND* project contribution. Retrieved from <http://open-mind.net/papers/why-and-how-does-consciousness-seem-the-way-it-seems>
- Friston K. (2009). The free-energy principle: A rough guide to the brain? *Trends Cognitive Science*, 13, 293–301.
- Friston, K. (2013). Active inference and free energy. *Behavioral and Brain Sciences*, 36, 212–213.
- Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 290, 181–197.
- Hinton, G. E. (2007). To recognize shapes, first learn to generate images. In P. Cisek, T. Drew, & J. Kalaska (Eds.), *Computational neuroscience: Theoretical insights into brain function* (pp. 535–547). Amsterdam, The Netherlands: Elsevier.
- Hohwy, J. (2007). Functional integration and the mind. *Synthese*, 159, 315–328.
- Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, 3, 96. <http://dx.doi.org/10.3389/fpsyg.2012.00096>
- Hohwy, J. (2013). *The predictive mind*. New York, NY: Oxford University Press.

- Hume, D. (1751/1998). *An enquiry concerning the principles of morals* (Tom L. Beauchamp, Ed.). Oxford, UK: Oxford University Press.
- Kail, P. (2007). *Projection and realism in Hume's philosophy*. Oxford, UK: Oxford University Press.
- Kveraga, K., Ghuman, A., & Bar, M. (2007) Top-down predictions in the cognitive brain. *Brain and Cognition*, 65, 145–168.
- Lupyan, G., & Clark, A. (2015). Words and the world: Predictive coding and the language-perception-cognition interface. *Current Directions in Psychological Science*, 24(4), 279–284.
- Pezzulo, G. (2014). Why do you fear the bogeyman? An embodied *predictive coding model* of perceptual inference. *Cognitive, Affective, and Behavioral Neuroscience*, 14(3), 902–911.
- Prinz, J. (2012). *The conscious brain: How attention engenders experience*. New York, NY: Oxford University Press.
- Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17, 565–573. doi:10.1016/j.tics.2013.09.007
- Zawidzki, T. (2008). The function of folk psychology: Mind reading or mind shaping? *Philosophical Explorations*, 11, 193–210.

## 7.2 REFLECTIONS ON ANDY CLARK

Daniel C. Dennett

Andy Clark has outdone himself as usual, to quote a fine old Ray Smullyan line. Our large domain of agreement has always been clear, and this time he doesn't just do justice to my project; he advances it in ways I had been struggling to get into focus. First, he expresses its goal more succinctly than I have managed:

It aims to show why we think experience presents a world populated with all that odd stuff (sweet tastes and cute babies) when actually, what we are tracking are just more features of the world—in this case, ones invisibly intertwined with our own reactive dispositions. (p. 206)

I wish I'd thought of putting it that way. Noting that I “must somehow deliver the apparent effects of qualia (the judgments we make, the things we say) without positing qualia as their cause or justification,” he then deftly unpacks the “Bayesian layer” of my argument, and he sees just how it would work:

Brought together under the accommodating umbrella of structured Bayesian predictive models, these all emerge as content-items (latent variables) inferred as part of the generative model that best predicts the raw sensory flow. Crucially, that sensory flow now includes a flow of predicted interoceptive information and response, and it is those aspects of the flow (reflecting our own reactive dispositions) that are responsible, if Dennett is right, for much of the apparent puzzlingness of qualitative experience. (p. 216)

Again, I wish I'd thought of putting it that way, and will in the future, also helping myself to his apt examples: Gregory's hollow face illusion, and the number-letter cross, vivid demonstrations of the swift power of "priors," both ancient and temporally local, to generate best explanations of the current sensory flow.

I could continue in this vein, quoting line after line with approval and gratitude (my annotated file of his chapter is ablaze with yellow highlighting), but let me leap instead to Clark's puzzle of "qualia surprise" and the strange case of the albino squirrel. How *could* we have qualia surprises in a Bayesian world? Happily, Clark solves his own puzzle:

The reason we can experience qualia surprise is because the raw sensory inputs here actually possess plenty of hidden structure—so spotting a novel kind of cuteness is really no harder than spotting a novel kind of chair as a chair or some crazy, futuristic automobile as an automobile. This suggests that it is our systematically structured grip on the world, realized courtesy of the neural prediction engine, that does the real work here. (p. 214)

This hidden structure is what our Bayesian brains first suss out and then conceal from "us" with a user illusion, allowing us to carry on with the important, time-pressured work of figuring out what to expect next. If "we" try "from the inside" (introspectively, reflectively) to figure out what is going on, we get the blur of all the "intrinsic," "ineffable" properties that seem to be the proximal causes of our unreflective beliefs about what is happening. As Clark puts it,

But (being the advanced, reflective agents we are) we may then diagnosis ourselves as having detected something that is deeply puzzling, given our larger view of the world. For qualitative cuteness now seems to have been "given" in the input stream, just like ordinary content properties such as baby-ness (or numeral 6-ness. If this is correct, then all those apparent cuteness-qualia (and all other qualia, including redness, painfulness, etc.) are really just disguised predictions of our own web of reactive dispositions. But they are predictions that, when satisfied, can appear to add mysterious qualitative dimensions to our daily encounters with a structured external world. (p. 212–213)

Compare Clark's cute albino squirrel to Hume's missing shade of blue. Hume was flummoxed by his unshakeable conviction that "simple" color "impressions" had to be what they seemed to be: atomic and unanalyzable, but his brute ability to fill in the gap in the series, effortlessly and indeed involuntarily, showed that there

was background machinery churning away, hinging on structure that he had no personal-level Need to Know. (Only nosy, inquisitive philosophers think they Need to Know such things, asking too many questions and getting themselves into all sorts of metaphysical difficulties. As Clark notes, “Such second-order predictions, by being borne out in practice, may indeed serve to cement our conviction that we are here dealing with real features of the world, by further obscuring the importance of our own reactive dispositions.”)

I arrive finally at Clark’s good closing question: “What remains of the claim that qualia are in some sense unreal?” As he notes, my view (Dennett, 2013) allies them with dogs and hats and hurricanes, and surely (ding!) these are as real as real can be. He suggests that a “modest” form of qualia-realism is available to me, minus the hypnotically seductive idea of double transduction, and he is right. What appears from most philosophical perspectives to be a deep metaphysical question about the Ultimate Structure of Reality can also be seen—more profitably, in my opinion—as a question of diplomacy or pedagogy, as I have sometimes put it. Should we say, “Xs don’t exist!” or “Xs are real; they just aren’t what you think they are.” With this big boost from Clark, I think we can “start to unravel the metaphysics of the Bayesian brain.”

## Work Cited

Dennett, D. C. (2013). *Intuition pumps and other tools for thinking*. New York, NY: Norton.







# EVOLUTION, SOCIALITY, AND AGENCY



# 8.1 TOWERS AND TREES IN COGNITIVE EVOLUTION

Peter Godfrey-Smith

What are the main stages in the evolution of cognition, the crucial transitions between simpler and more complex forms? When addressing this question, it is tempting to adopt a kind of linear thinking— $X$  leads to  $Y$  which leads to  $Z$ —and common also to push back against these temptations. Evolution is not a matter of ladders and scales, we’re reminded, but a process of branching and divergence. Discrete stages, as opposed to seamless gradations, might also be seen as unlikely. But amid all the radiations and gradual shifts, and all the different niches and lifestyles, might there be some especially important inventions that fall into a natural order? Perhaps such a sequence might be important, also, as a manifestation of some general principle about how organisms control behavior and deal with the world.

One of my favorite papers by Dan Dennett is “Why the Law of Effect Will Not Go Away” (1975; reprinted in *Brainstorms*, 1978). Here Dennett made some early moves (following up the even earlier *Content and Consciousness*, 1969) in what has become a lengthy defense of the idea that there is a single pattern essential to all processes of adaptation, cognitive improvement, and “R & D” in a broad sense of the term. That pattern is the Darwinian one: trial and error, generate and test, variation and selection. Dennett argues that whenever adaptive improvement occurs, a generate-and-test process must be at the bottom of it. Such a process can be realized on many time-scales; Thorndike’s “law of effect” of 1905 is a manifestation of the pattern within an individual learner’s lifetime. Successful behaviors are repeated, unsuccessful ones abandoned.<sup>1</sup>

---

1. The idea that variation and selection are omnipresent in adaptive processes was also defended around the same time by Donald Campbell (1974), and one kind of “evolutionary epistemology” takes off from this idea.

In his books *Darwin's Dangerous Idea* (1995) and *Kinds of Minds* (1996), Dennett developed these sketches into a "Tower of Generate and Test," with a series of transitions between ways of realizing the Darwinian pattern on different scales and with different degrees of sophistication. At the bottom are *Darwinian creatures*, which can only adapt through genetic mutation and natural selection. Above them in the tower, *Skinnerian creatures* also adapt during their lifetimes by trial and error learning. Above those are *Popperian creatures*, which don't have to actually perform the behavioral options being tested. Instead, they can run internal experiments to assess the consequences of a behavior "offline," discarding bad options and only exposing the better ones. *Gregorian creatures* use social tools, especially language, to make use of innovations discovered by other individuals. Rather than test all the options themselves, even internally, they draw on the collective experience of many.

In 2015, at a conference at Macquarie University, the psychologists Russell Gray and Alex Taylor discussed a more elaborate tower, due to Taylor, that added a number of floors to Dennett's.<sup>2</sup> But both Gray and Taylor worried that a "tower" is not the right framework to be using at all; it has too much of the *scala natura* about it. Darwinians don't expect scales but trees. That concern is reasonable, but there can still be cumulative and directional processes in a Darwinian context, processes in which one move must occur before another. So given all that we now know in comparative psychology and evolutionary biology, what should we make of Dennett's tower and other attempts to mark out the stages in an advance of cognitive complexity?

I'll investigate this question by offering a new taxonomy of creatures, drawing on Dennett, Taylor, and other work. Dennett's own tower had a dual role. First, it was intended to mark out genuine historical stages, though it was not intended as an overall map of the evolution of the mind, which involves a lot more than the proliferation of generate-and-test mechanisms. (Perhaps Dennett is not so convinced there's a *lot* more.) Second, Dennett sees the sequence of realizations of the Darwinian pattern as part of a general account of how adaptation, design, and meaning are possible in a wholly physical world. My project here lies more in the history of cognition, and I will set up a sequence in which generate-and-test mechanisms do have an important role but are mixed in with others. My interest is partly in the question of where there *is* a tower-like structure, and where there is not, both empirically and in principle. By "tower-like" structure I mean a situation in which the evolution of one cognitive mechanism must, or at least always does, precede the

---

2. The conference was *Understanding Complex Animal Cognition: An Interdisciplinary Workshop*, organized by Rachael Brown, at Macquarie University, Sydney, Australia.

evolution of another, especially in situations where both are retained. If this situation holds, there will be a nesting: every organism with *Y* will also have *X*, and so on. For which cognitive kinds is there such a sequence, and for which is there a different pattern?<sup>3</sup>

If there is nesting in some cases, this will still be laid out as a tree in the basic sense relevant to Darwinism. You might have a situation where *X* appears early and *Y* later appears independently in a subset of the branches with *X*. Then not all the branches will have the same things on them, but if we set aside losses, there will be no cases where *Y* appears without *X*. When that is true, there is something of a tower present within the Darwinian tree.

Before going on, I'll note a terminological point. A categorization like the one seen here can be set up either with nested or exclusive categories. (Are Skinnerian creatures *also* Darwinian creatures, or does becoming Skinnerian take an animal out of the Darwinian category?) Dennett has mostly (perhaps not always) written about his categories as if they are exclusive. Given the way my story will turn out, it's best to use nested and non-exclusive categories (so a Skinnerian can also be Darwinian). Sometimes I'll say "*merely* Darwinian," or something similar, to indicate contrasts, and occasionally this role will be played by the context, in a way that I hope remains clear.

## 1. Humean Creatures

In Dennett's tower, there are Darwinian creatures, which evolve by natural selection over generations, and then Skinnerian creatures, which adapt during their lifetime by trial and error learning. I'll discuss a stage between these—though whether it really lies *between* them will be discussed in detail. These are *Humean creatures*, creatures capable of associative learning that does not have a trial-and-error character, also known as classical conditioning.

I take the essential feature of classical conditioning to be the learning of correlations between events perceived. Pavlov's dogs regularly heard the bell ring before the arrival of food. The usefulness of classical conditioning is predictive; you can learn that one event predicts another. There's an *unconditioned stimulus* (US), to which you have a preexisting behavior, an *unconditioned response* (UR), that has been established through evolution or earlier learning. You learn

---

3. There might be a directional process in which replacement rather than retention was the prevailing pattern—*X* is always *replaced* by *Y*, and so on. I'll assume in this chapter that capacities are augmented rather than replaced, and will ignore the possibility of losses except where this is explicitly discussed.

to produce this behavior in response to a new *conditioned stimulus* (CS) which is found to be associated with the US.

I used Pavlov's example but named the creatures after Hume. I do see Hume as the first to describe a mechanism of this general kind. As Hume had it, ideas that occur in the mind together tend to prompt or stimulate each other. A kind of "attraction" operates between them. There is no essential link to behavior or reward in Hume's story. Though he was attuned to adaptive considerations, in a loose sort of way, Hume conceived of his mechanism more along the lines of a quasi-physics, a "power of attraction" in a sense analogous to Newtonian gravity. After Hume, this picture evolved into a psychological theory, via James Mill and his son J. S., also Bain and Spencer, and then the experimental tradition of Pavlov, Watson, and others.<sup>4</sup> People sometimes talk of "associative learning" as a single thing, or a single package with minor variants, but I am emphasizing what I take to be some deep differences between its forms.

Who is a Humean creature and how did this capacity evolve? Classical conditioning is *very* widespread. To put the distribution of Humean capacities into context, I will give a quick sketch of the evolutionary history of animals. Animals evolved from single-celled protists somewhere between 700 million to a billion years ago.<sup>5</sup> The early branchings are currently unclear. It used to be thought that sponges are a "sister group" to all other animals—they are the present-day animals whose lineage branched off first. Some data presently suggest that *ctenophores* (comb jellies) branched off earlier. If so, there may have been two independent origins of nervous systems, as ctenophores have them and sponges do not. Either way, later than the branching between sponges and others was a split between cnidarians and a large group of *bilaterian* animals, those with a left and right side as well as a distinction between top and bottom. Nearly all the familiar animals are in this group. Cnidarians, which are radially rather than bilaterally symmetrical, include jellyfish, anemones, and corals. Within the bilaterians is a further split into *protostomes* and *deuterostomes*. Protostomes include arthropods (like insects and crabs), molluscs (like clams and octopuses), annelids (like earthworms) and some others. In deuterostomes, the main groups include vertebrates, like ourselves, and echinoderms, including starfish.

I will draw on a 2013 review of invertebrate learning by Clint Perry, Andrew Barron, and Ken Cheng (2013), which includes a large chart summarizing which animals have been shown to do various kinds of learning. Classical conditioning is extremely widespread among bilaterians, and it is seen in some very simple

---

4. For this history, see Greenwood (2009).

5. See Budd and Jensen (2015); Peterson, Cotton, Gehling, and Pisani (2008); Ryan et al. (2013).

animals.<sup>6</sup> The nematode worm *Caenorhabditis elegans*, which has only 959 cells and 302 neurons (in the hermaphrodite) has shown classical conditioning. In the summary given in the Perry et al. (2013) paper, there are only a few bilaterian groups in which classical conditioning has *not* been shown, including millipedes, barnacles, rotifers, and also sea squirts, which are quite close to us on the tree but lead simple lives.

Classical conditioning might, then, be an early bilaterian invention, evolving once and ramifying through the many groups, perhaps lost in some, or just hard to find. If this is true, classical conditioning is very old—600 million years or so. Alternatively, it might have been invented several times, being useful and easy to build. I've not been able to find much information, or even confident claims, on that last point. Classical conditioning has also been reported in one animal (in a study accepted by Perry et al. [2013]) outside the bilaterians: in an anemone.<sup>7</sup> So if classical conditioning has a single origin, then it is even older than I just said it is, or perhaps there's one origin here and another (or several) in the bilaterians. Some old reports suggest classical conditioning in paramecia, which are single-celled protists, and not animals at all. But these reports seem to have been discredited, as far as I can tell, and I will set those aside.

The Perry, Barron, and Cheng (2013) paper acknowledges the difficulty in showing that an animal *can't* do some form of learning. If the experiment was done at all, it might have used the wrong stimuli; or, it might not have been reported because negative results are seen as less interesting and harder to publish. Their paper is a list of what has been shown to be present, not what has shown to be absent. In some cases here, however, I will take—very provisionally—the absence of evidence to be evidence of absence. In well-studied organisms, the non-report of an interesting trait is informative. At other places I will treat absences more hypothetically: *if* we assume the distribution is a certain way (including absences), then a certain picture results. Anyway, a very large range of animals are Humean creatures, and there even seems to be a radially symmetrical Humean.<sup>8</sup>

The evolutionary history of classical conditioning also bears on other questions. In the early evolution of animals there is an enigmatic period before the Cambrian explosion, called the *Ediacaran*, between about 635 and 540 million years ago. Before this time there is no fossil trace of animal life at all. In the Ediacaran there are fossils of many sizes and kinds. It's hard to work out how these animals lived and, in many cases, whether they were animals at all. Some have

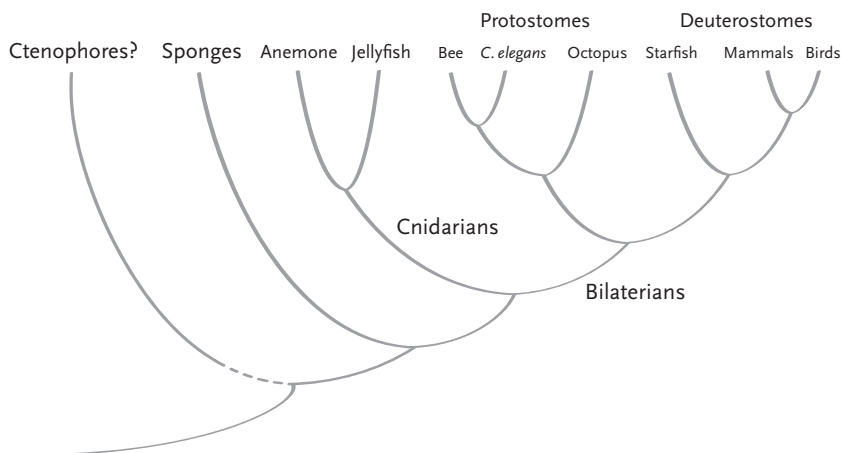
---

6. The main figure in the paper does have one typographical error: the molluscs are split in a way that is not accurate.

7. See J. Haralson, Groff, and S. Haralson. (1975).

8. Cubozoa are usually seen as the most behaviorally sophisticated non-bilaterians. I don't know if anyone has looked there.





**FIGURE 1.** Evolutionary relationships between some major animal groups. The figure is not to scale with respect to time, and it mixes taxonomic ranks. The order of the early branchings is controversial.

three-sided forms that have never been seen since. Genetic evidence strongly suggests that animals were around, and that they had evolved nervous systems, even before this time. The branching between cnidarians and bilaterians in Figure 1 may date to 650 or 700 million years ago, and the animals on both sides have nervous systems.<sup>9</sup> Behavior in the Ediacaran seems to have been very simple, though; there are no known bodies with legs, claws, sophisticated eyes, spines, or shells. These animals seem to have lived lives much more self-contained than a typical animal life now. This limited traffic with the world makes Ediacaran animals appear to be—in a phrase suggested by Hans Pols—*Leibnizian creatures*. They were not truly windowless—no life is windowless, even the simplest known bacteria have sensory “windows” (signal transduction mechanisms) with which they track what is outside.<sup>10</sup> But the evolution of animals might have included a somewhat less windowy time. This picture becomes less likely, however, if classical conditioning was around. Classical conditioning is essentially a tool for dealing with external patterns and events, in situations where timing matters. So *if* it were to be shown that classical conditioning is of Ediacaran age, this would suggest that this time was not so Leibnizian—unless the only living remains of that time are descended from the animals most interested in looking out the window.<sup>11</sup>

9. See Peterson et al. (2008).

10. See Lyon (2015).

11. In standard taxonomies of learning, there is also “non-associative” learning—habituation and sensitization. In the Perry et al. (2013) review, habituation is represented as even more

## 2. Skinnerian Creatures

Skinnerian creatures, in Dennett's hierarchy, learn by trial and error, tracking the good and bad consequences of their actions. They retain and reproduce actions that lead to good outcomes and discard ones that don't. I'll refer to this broad category as *instrumental* learning or conditioning.<sup>12</sup>

Instrumental learning features an essential role for reward and/or punishment. It's not merely reward-driven behavior in an immediate, here-and-now sense (pull back when shocked, keep coming when rewarded). It has to involve longer-term change and the shaping of dispositions, so the animal responds adaptively to the *next* situation of a certain kind, as opposed to maintaining or ceasing a behavior in response to what it experiences in the present.

Drawing again on the review by Perry, Barron, and Cheng (2013), who can learn in this way? First, it is found only in bilaterian animals. It has also not been found in nearly as many bilaterians as classical conditioning; there are many gaps. It is seen in some arthropods (bees, crabs, crickets, and some others), but there are others in which it has not been reported (spiders, wasps, millipedes). Of those, spiders and wasps can be classically conditioned, though it has not been reported in millipedes. In molluscs, some can, and some (apparently) can't. No worms have been reported to do it, though annelid worms can be classically conditioned.

If we take this survey at face value, animals who can learn by instrumental conditioning are a strict subset of those who can learn by classical conditioning. That is: *all Skinnerian creatures are Humean creatures, but not vice versa*. If so, we have a cumulative, tower-like structure so far.

The distribution of this trait also makes it seem likely that instrumental conditioning did not evolve once and get passed down many branches of the tree. It's likely to have evolved a number of times. Otherwise, instrumental learning has very often been lost, and/or many non-reports of the trait are misleading.

I don't discount that last possibility, and there are interesting grey areas, too. Nematode worms are an important case in which classical conditioning has been shown, but, in the surveys I'm using, instrumental learning has not. These are

---

widespread—seen in jellyfish and anemones, for example, as well as all the bilaterian groups they include. Sensitization is not quite as common. These forms of plasticity are so minimal I leave them out of further discussion, though. It is true that there are interesting possibilities concerning the evolution of associative from non-associative learning (Wells, 1968), and there is also a detailed and interesting report of habituation in plants (Gagliano, Renton, Depczynski, & Mancuso, 2014).

12. It is sometimes called *operant* conditioning, but I'll use that term in a narrower sense (as some others do). In this sense, operant conditioning involves the addition of new behaviors to a repertoire, rather than merely adjusting the frequency or setting in which pre-existing and stereotypical behaviors are produced. The Perry et al. (2013) review uses a broader sense of "operant."

very well-studied animals. A closer look at the literature shows some borderline phenomena, too. Although a lot of what nematodes can do fits classical conditioning, they have been reported to show rapid aversion learning by taste. Food that makes them sick is avoided when it appears again, in a way that has been compared to the “Garcia effect” in mammals.<sup>13</sup> Might this, also, be a Humean phenomenon? Perhaps the taste of the (bacterial) food is associated with an experience of unease that has an unconditioned response (UR) of avoidance. Then might the taste come to be used as a predictor of that experience, leading to a pairing of the avoidance with the taste? Perhaps, and in a moment I’ll look more closely at attempts to Humeanize behaviors of this general kind.

Another borderline case is further away on the tree: anemones again. They have not been reported to engage in instrumental learning, but in a recent study, Mark Briffa found the following. In an anemone species in which individuals sometimes fight each other, there are meeker and bolder anemones. Briffa found that anemones become less bold after they lose a fight, and the change in boldness is seen across different contexts, not merely in later fights. Is this a case of instrumental learning? Do they learn *that* they’re no good as fighters, or that fighting is a bad idea for them?<sup>14</sup>

Pressing on these borderline cases, especially the one with nematodes, leads to a general problem. What kind of separation is there, in principle, between classical and instrumental conditioning, between Humean and Skinnerian capacities? Many papers assume the distinction and treat it as an important classificatory tool. But quite a few discussions, especially by specialists, either reject the standard distinction or see it as very elusive in contexts of testing.<sup>15</sup> Björn Brembs, for example, has suggested that the distinction “needs to be reconsidered,” as most situations in which associative learning occurs can be described in either way, or perhaps contain both processes. When an animal performs some behavior, gets a reward, and consequently performs the behavior more often, this can be

---

13. See Ardiel and Rankin (2010); and Nuttley, Atkinson-Leadbetter, and van der Kooy (2002). For the “Garcia effect,” see Garcia and Koelling (1966).

14. Rudin and Briffa (2012). Limpets, a kind of gastropod not on the Perry et al. (2013) list for instrumental conditioning, have been reported to do something similar: see Shanks (2002). Spiders have also been reported to do something like this—fight winners (and losers) behave differently the next time they fight—but the decay of memory is quick, with a complete reset by 24 hours (Kasumovic, Elias, Sivalinghem, Mason, & Andrade, 2010).

15. The nematode case illustrates the problem. Andrew Barron, one of the authors of the Perry et al. (2013) review I use, is very cautious about these categorizations (personal communication). For the Brembs comment, see Björn Brembs, (2013, July 11), Brains as output/input systems [Web blog post]. Retrieved from <http://bjoern.brembs.net/2013/07/brains-as-outputinput-systems/>.

described in terms of an association between behavior and reward, *or* in terms of a classically conditioned association between the experience of performing the behavior and the appearance of (say) food. The animal learns *that* the experience of doing *X* predicts the arrival of food, much as it might learn that a bell predicts food.<sup>16</sup> Brems thinks that with care it is possible to remove the classical element from some cases of instrumental learning, but it's not easy, and the results should then be seen as akin to the learning of skills.

In this way and others, the simple and neat Humean/Skinnerian relationship could be made more complicated. I'm reluctant to admit that the distinction is problematic in principle, though. Consider some extreme cases. An animal could be completely blind to reward—completely unable to distinguish between good and bad outcomes—and yet be able to learn by classical conditioning; it could still learn that when *X* happens, *Y* happens soon after, and it could come to produce its *Y*-adapted behavior in response to *X*, as well. It might not register the good consequences of that production of a *Y*-adapted behavior; only blind Darwinian processes might assign the appropriate behavior to stimulus *Y*. But it can usefully modify when the behavior is produced. An animal might be sensitive to facts but not to the consequences of its actions and still usefully learn.

On the other side, can an animal be sensitive to the consequences of its actions but not to the facts? In a sense, no, because the consequences of actions *are* facts. An animal sensitive to consequences can't be insensitive to all the facts. (This is part of what Brems was getting at.) However, whereas a sensitivity to facts, not consequences, can be sufficient to change behavior using classical conditioning, the kind of "facts" discussed just above (facts about the consequences of actions) can't be used to change behavior in that way. It might be a fact that lever-pressing predicts the arrival of food, but lever-pressing only occurs when the animal decides to *do* it. It's not an exogenous event that can be sensed and then associated with other events. The animal might learn the fact that next time it does a lever press, food will arrive, but in a classical conditioning scenario, that fact cannot induce the animal to do the lever press. The animal would have to wait until presses occurred, at which point it could expect food and act in a way adapted to food's arrival. But "waiting" is not a way for the lever to become pressed, at least in normal circumstances. Something must induce the animal to *make* the press.

In instrumental conditioning, the initial presses might be entirely random, and the later ones are guided by reward. Pressing occurs *because* it has had good effects. If an animal is only classically conditionable, there is no reason why it should go from an initial random lever press to additional ones. All it can do is

---

16. Dickinson and Balleine (1994).

note that *if* the lever happens to be pressed, there are other things it can expect to happen.

If sensitivity to correlations between perceived events and sensitivity to reward are two distinct capacities, then we can ask straightforward questions about which comes first, if either does, and which animals have one without the other. The minimally sufficient mechanisms for each kind of learning are different, too. Simple arrangements of Hebbian synapses are sufficient for classical conditioning: suppose neurons *A* and *B* both excite *C*, with *A* but not *B* being initially sufficient to make *C* fire. Then if *A* and *B* register distinct but correlated events, *B* will fire when *C* does (as *C* has been made to fire by *A*), and the link between *B* and *C* can strengthen according to Hebb's rule (neurons that fire together, wire together), until *A* is also sufficient to make *C* fire. This is a local and undemanding process. Nothing so simple suffices for instrumental conditioning. The animal must have a way of registering the consequences of its actions in a way that feeds back and affects how it will behave in the future. The apparent distribution of Humean and Skinnerian capacities on the phylogenetic tree appears to reflect this difference in how demanding the processes are. From what I've said, there's no reason why an animal *couldn't* have instrumental without classical conditioning. But perhaps classical conditioning is easy enough and useful enough for this never to happen, whereas instrumental conditioning—which is even more useful, one might think—is not so easy.

It might reasonably be objected that I am assuming here a very simple form of classical conditioning—a form that is simpler than what many animals seem to use. Animals don't just track where two stimuli occur near to each other in time; they track whether the conditioned stimulus is genuinely *predictive* of the unconditioned stimulus. This is a much harder question for the learning brain to answer, but classical conditioning seems to be sensitive to it.<sup>17</sup> And this is the tip of the iceberg. Here is a result discussed in this connection by Anthony Dickinson, from an experiment by Balleine, Espinet, and Gonzalez (2005). Rats were fed initially on a sweetened drink that had two flavors, orange and lime, combined. Then they let the rats drink an unsweetened lime-only drink. Later, those rats showed a stronger preference for an orange-only drink than did rats for whom the second lime-only drink *had* been sweet. Put in folk-psychological terms, the rats who drank the unsweetened lime-only drink worked out that it was probably the orange that was the source of the sugar when they'd encountered the two together. Tasting the unsweetened lime-only drink made them *retrospectively re-evaluate* the orange flavor.

---

17. Rescorla (1967); Dickinson (2012).

Contemporary work on associationism posits mechanisms that have much more complexity than Hume or Pavlov envisaged.<sup>18</sup> There also seems to be a tendency, though, to try to give quite *unified* models of how classical conditioning and instrumental conditioning work.<sup>19</sup> But if classical conditioning arose independently more than once, there's no reason to expect it to work the same way all the time. It might appear in simpler forms in some animals and complicated forms in others. Even if it arose just once, it presumably arose first in a simple form, and there would then be no reason to expect it to become complex in the same way in all the independently evolving lineages on which it's now found. Or perhaps there *is* reason to expect this, if there's a single way of doing it that is both more complex than the initial form and the best possible way (or a very good and easily engineered way) of doing it. Dickinson (2012) does not consider multiple evolutionary lineages, but he does consider such a sequence: "The ancestral form of associative learning may have been based on simple temporal contiguity between events. However, this simple system was prone to developing superstitious 'beliefs' based on fortuitous event pairings" (p. 2737). From there, a more complex process arose that made the organism sensitive to the predictive relationships, not merely a pairing, between events.

As Dickinson sketches it, the simple version evolves, and then the complex version follows. But given the shape of the evolutionary tree, there are many possibilities. Perhaps the simple version evolves once, and the same complex version appears in various places later on because it's so useful. Perhaps the same path from nothing to simple to complex is trodden many times. Or perhaps both the simple and the complex forms were one-time inventions, much deeper in the past than a paleontological behavioral ecologist would ever have suspected. It's interesting, given this sketch of possibilities, that the old anemone study I mentioned earlier

---

18. Dickinson (2012):

Within contemporary associationism, knowledge takes the form of what is often called a representation of the relationship between events, be they stimuli or responses, by a connection (or association) between representations of these events. The process by which this knowledge is deployed is the transmission of excitation or activation (and inhibition) from one event representation to the other via the connection with stronger connections producing greater transmission. Finally, this associative knowledge is acquired by the progressive strengthening of the connection with each effective experience of a relationship between the events. (p. 2733)

Associationism is no longer a quasi-physics, or does some of that character remain?

19. See, for example, the way the cricket work in Terao, Matsumoto, and Mizunami (2015) is set up.

claims to have shown a complex form of classical conditioning, one attuned to predictive relationships. And a recent study of crickets reported that they, too, exhibited a quite complex form of classical conditioning.<sup>20</sup>

I finish this section with a point of purely historical interest about instrumental learning and its relatives. Dennett used Skinner as namesake for this kind of creature. Why not Thorndike, as in Dennett's old paper? I associated Hume, an 18th-century name, with classical conditioning, so we might wonder whether an earlier figure deserves the credit here as well. In intellectual history, as well as animal evolution, instrumental learning seems to have come later. One of the first to describe it was Alexander Bain, in his 1859(!) book *Emotions and the Will*.<sup>21</sup> Bain's work, a mix of psychology and philosophy, is quite neglected, though the American pragmatists recognized its importance. I suspect that Dennett gave Skinner the credit because Skinner saw and emphasized the analogy between instrumental learning and Darwinian evolution (1938). Bain did not; and I don't know whether Thorndike did. William James saw it, and James taught Thorndike, but James tended toward saltationism about both kinds of evolution in a way that alters the explanatory role of selection quite substantially.<sup>22</sup> Hume (1739/2000), incidentally, did hit on a momentary anticipation of adaptation by trial and error, in a passage about social behavior:

Two men who pull the oars of a boat, do it by an agreement or convention, although they have never given promises to each other. Nor is the rule concerning the stability of possessions the less derived from human conventions, that it arises gradually, and acquires force by a slow progression *and by our repeated experience of the inconveniences of transgressing it* [italics added]. (bk. 3, pt. 2, section 2)<sup>23</sup>

---

20. The paper by Haralson et al. (1975) reports that the work did include a Rescorla-type "truly random" control. The cricket study is Terao et al. (2015).

21. Bain (1859): "In the primitive aspect of volition, which also continues to be exemplified through the whole of life, an action once begun by spontaneous accident is maintained, when it sensibly alleviates a pain, or nurses a pleasure." I had thought Bain might deserve credit for being *the* first to state a principle of this kind, but according to John Greenwood (2009), he was not the first and he might have picked up this idea from the German psychologist Johannes Müller (1801–1858).

22. See Godfrey-Smith (1996, chapter 3).

23. Another near-miss is monumentally ironic. Lamarck, in his 1809 defense of his doctrine of the inheritance of acquired characteristics, notes that a version of his view exists as a proverb, "*Habits form a second nature*." Then, "if the habits and nature of each animal could never vary, the proverb would have been false and would not have come into existence, nor been preserved in the event of anyone suggesting it" (1809/2011, p. 114).



It is interesting that Hume's theory of psychological dynamics did not include anything like this, another aspect of the blind spot about selectionist mechanisms that seemed firmly in place before Darwin and Bain.

### 3. Carnapian, Pearlian, Mimetic, Popperian, and Tolmanian Creatures

Dennett's tower had *Popperian creatures* next. Rather than trying out behaviors in potentially risky actions, they internalize the generate-and-test mechanism entirely. I'll work my way to these creatures but will do that by thinking more generally about steps that might be made onward from the Skinnerian state.

At this point it's surely likely that the tower-like structure breaks down. I'll discuss four separate sophistications that might appear next, without offering an ordering of them. I'll also suggest that the Popperian grade of cognition builds on elements from this set—perhaps one, perhaps several.

The first thing to mention is not a step to a new algorithm or mechanism but a road forward that can be trod to different degrees. In current work on learning in some invertebrates, especially bees, there is a lot of interest in the learning of abstract concepts and logical relations—conjunctions and disjunctions, relational concepts such as *larger than*, and others. In the review by Perry, Barron, and Cheng (2013) used in earlier sections of this paper, this capacity to deal with abstraction is the main path away from ordinary instrumental conditioning that they consider (along with navigation, which I'll discuss below). Bees seem to be the most notable masters of abstraction among the invertebrates—or at least, the animals in which abstraction has most clearly been demonstrated so far. For example, after being conditioned to choose the *larger* of two stimuli, bees can extrapolate this rule to a situation in which two objects of different sizes from either of those they saw during training are presented to them, and in which the object most similar to the object they learned to choose during the training period is *not* the one they need to pick in the new situation, because it's not the larger of the new pair.<sup>24</sup> That is not bad for a brain one cubic millimeter in size. A likely empirical picture, in any case, is that one road away from simple instrumental learning is a road on which the basic rules are the same, but the concepts and discriminations employed are more abstract and complex. I'll call animals like bees *Carnapian creatures*.

---

24. Avarguès-Weber, Dyer, Combe, and Giurfa (2012); Avarguès-Weber, d'Amaro, Metzler, and Dyer (2014).



Here is a second kind of sophistication, which might be reached by traveling the Carnapian path, but might not be. Taking up a concept from Taylor's tower, I'll recognize *Pearlian creatures*. These animals can make use of causal reasoning based on a notion of intervention, the kind of causal thinking described in detail by the computer scientist Judea Pearl and also by a number of philosophers, including Jim Woodward, Clark Glymour, Richard Scheines, and Peter Spirtes. The psychological application of these ideas has been developed by Alison Gopnik.<sup>25</sup> The controversial idea to look at here is that thinking based on a rich notion of cause is different from thinking based on mere extraction and extrapolation of patterns. That is, Carnapian and Pearlian creatures might not coincide, and one might go far down the Carnapian road without becoming a Pearlian creature, or vice versa.

Here is an example used by Taylor to make a point like this (drawing on Tomasello and Call, 1997).<sup>26</sup> Suppose you see the wind move a tree branch, and some fruit falls. Humean creatures might learn an association between wind and fallen fruit, which could be useful. The next time they feel wind, they expect to find fruit. The extraction of patterns from what is seen might be very sophisticated (they might track how much wind tends to predict which particular kinds of fruit will fall, and so on). But a different sort of creature could see the fruit fall and realize, Aha! I can just move the tree branch myself! No need to wait for the wind.

If you just happened to *do* a tree shake and fruit resulted, you could succeed in this context as a Skinnerian creature. You could learn to go around shaking trees. That requires that you have some reason to perform the action the first time. As a Skinnerian, you might do things of this kind at random. But that is not very efficient, and a Pearlian creature does not need to do this. From the experience of seeing the wind shaking the tree, the animal can work out that the same effect can be achieved by an intervention. The creature who does this might not be thinking very abstractly—might not be far down the Carnapian road—but is thinking causally, and that's a powerful thing.

Which animals are Pearlian creatures? Taylor, working with a team of bird people and causal reasoning people (2014), did an experiment with New Caledonian crows that was designed to probe this question, and he did not find Pearlian behaviors. In contrast, two-year-old children passed the test. Their experiment set up a situation analogous to the wind–fruit scenario. The crows were trained on a puzzle box in a way that enabled them to sometimes get food but in a way that

---

25. See Pearl (2000); Spirtes et al. (2000); Gopnik and Schulz (2007); Woodward (2005),

26. See Tomasello and Call (1997).

should have made it clear to a Pearlian creature that some novel manipulations (akin to shaking the branch) were available. Children picked up on the shortcuts and crows did not. The crows, in contrast, seemed able to approach the problem with instrumental conditioning—rewarded behaviors were repeated—but that’s all they could manage.<sup>27</sup>

I’ve distinguished two ways of becoming more sophisticated than an ordinary Skinnerian creature, the Carnapian and the Pearlian roads. Perhaps some facility with abstraction is needed to become a Pearlian, perhaps not. Either way, I suggest that once an animal has the Pearlian capacity, there’s a natural path to becoming a *Popperian creature* in Dennett’s sense. The crucial step to this sort of intelligence is the ability to run an experiment involving action off-line, to work out what would happen if I were to do *X*. If an animal can choose the most efficacious action in this way without performing it first, then the trial-and-error mechanism of Darwin and Skinner has been entirely internalized. As Popper said, this capacity enables our theories to “die in our stead.”<sup>28</sup> Not only need the Popperian organism not die for its experiments, but it need not encounter aversive experiences, either. The aversive experiences need only be envisaged.

Being a Popperian creature in this sense is quite demanding. People are such creatures. I take it that it’s not completely clear that any other animals are, though some mammals and birds are possibilities. Dennett, in some of his discussions, sets a much lower bar for being a Popperian creature. He says that you’re a Popperian as long as you don’t emit new behaviors completely blindly, but use information from the environment to filter which actions are chosen and exposed to the perils of reinforcement. If this is the bar, then as Dennett says, fish, primates, and many other animals are Popperian. But here I think Dennett mishandles his own hierarchy. If the bar to clear in order to become Popperian is just that there be some experiential filtering of which behaviors are initially performed, then even classical conditioning suffices. In these discussions, Dennett’s “filter” on behaviors apparently need not be a consequence-assessor. But there *does* have to be an internal consequence-assessor in order for the mechanism to retain a variation-and-selection character. It’s probably quite difficult for an animal to fully internalize the “test” part of the generate-and-test mechanism, as opposed to just having *some* way of shaping its other actions other than reinforcement of actual performance. And having a separation between a generate phase and a test

---

27. Taylor et al. (2014). “Aesop’s fable” cases are also possibilities (Bird & Emery, 2009).

28. Popper said this in his Darwin lecture, “Natural Selection and the Emergence of the Mind,” November 8, 1977, Cambridge University, England.

phase, where the test phase is truly internalized, may require a capacity to run internal experiments using causal reasoning.<sup>29</sup>

An objection that might be raised at this point is that a Pearlian creature, in the sense I have described, *is* a Popperian one. There is no way to achieve a “grasp” of causal relations and put this knowledge to use in an intervention if not through a Popperian generate-and-test mechanism. Perhaps that’s true, but I leave open the possibility that there can be causal cognition of a substantial and definite kind without the offline experimentation characteristic of the Popperian. Another objection that might be made is that one can get to the Popperian mechanism without using a strong, intervention-based notion of causation, via a purely Carnapian road. Again, perhaps that’s true, but it might not be. I suspect that the ability to try out behavioral options offline has no obvious relation to abstraction—it might be done while only noting simple relations between objects. We see that in Taylor’s experiment with babies and crows.

I’ll make one more point about this trio of categories: Carnap, Pearl, Popper. Some readers might have wondered before now about Wolfgang Köhler’s (1924) notion of *insight*. Köhler, working in the early twentieth century, thought that some of the behaviors he saw in chimps showed a grasp of the causal properties of objects, and an ability to use those objects to achieve novel solutions to problems (such as stacking boxes to climb on and reach a high banana). The idea of insight has often been seen as suggestive but problematically vague. I think insight might best be viewed as some sort of combination of the things I’ve discussed so far in this section—causal cognition, abstraction, offline experimentation. It’s not a definite psychological kind—there are no “Köhlerian creatures”—but various Köhlerian behaviors arise in different ways from this family of mechanisms.<sup>30</sup>

I’ll now introduce another path to the roads beyond Skinner. All the mechanisms discussed in this section are individualistic. Another piece of the picture is social learning. In Dennett’s tower, *Gregorian creatures* are found above Popperians. They improve their generate-and-test abilities with socially based tools for thinking, such as language. But once social learning is put on the table, we need to consider it in its own right. Social learning is not something that arises “after” Popperian mechanisms, the next level of a tower. It’s more widespread, and reasonably seen as a distinct step away from purely Skinnerian creatures, with its

---

29. A literature that must bear on this—but whose exact bearing I’ve not yet been able to work out—is the literature on “counterfactual” thinking in nonhumans, including rats, and also a literature on the role of “fictive,” as opposed to actual, rewards gained by an animal. See, respectively, Laurent and Balleine (2015); and Kim et al. (2015). This is a topic for discussion on another occasion.

30. See Shettleworth (2012).

own role in behavioral adaptation. One thing an animal can do is make use of another's history of trial and error, taking on that animal's behavioral adaptations as its own.

The role of simple forms of social learning was illuminated by an elegant model due to Alan Rogers (1989). Assume a population in which everyone is learning a behavior by trial and error of a Skinnerian kind. You can economize on the process, and behave just as adaptively as the others, by copying the behavior that is widespread around you. A population of Skinnerian learners can be invaded by an imitating mutant. I'll call this a *mimetic creature*. These creatures, who blindly copy, can invade Skinnerians, who pay the cost of trial and error. But when the mimetic creatures become very common, their copying becomes less effective, as they are no longer reliably copying someone who really knows what to do—the model assumes that the environment changes from time to time, and with it the appropriate behavior changes, too. Eventually, Skinnerians will bounce back. They pay the cost of learning, but when they are rare, they are the only ones likely to be performing the right behaviors. An equilibrium will be reached. Rogers also showed that (given some other reasonable assumptions) the average fitness in the population when the Skinnerians and mimetics are at equilibrium is the same as it is when the population consists only of Skinnerians.

Recent work has shown the presence of imitation in some surprising cases—animals without notable social lives. Examples include the red-footed tortoise (Wilkinson, Kuenstner, Mueller, & Huber, 2010), a stingray (Thonhauser, Gutnick, Byrne, Kral, Burghardt, & Kuba, 2013), and more controversially, an octopus (Fiorito & Scotto, 1992). These findings are used, with other arguments, by Cecilia Heyes (2011) to argue that “social” learning is not as distinctive in its mechanisms as people often suppose; she thinks that in at least many cases, it arises from the use of ordinary associative processes in a social context. If that is true, setting up a distinct category of “mimetic” creatures would be a bit misleading.<sup>31</sup>

Distinctions can be made between simpler and more sophisticated forms of imitation—you can merely copy what is common or visible around you, or instead copy only what seems to be successfully employed by others. The evolutionary consequences of sophisticated forms of imitation can be enormous, as Michael Tomasello has argued in his account of human cognitive change (1999). The distribution of simple forms of imitation above shows that this is another trait that has almost certainly evolved independently several times. That's undoubtedly true if the octopus results stand up but also likely given the presence

---

31. She does think that social learning may often involve specializations on the input side—some animals pay special attention to the behavior of others.

of this behavior in stingrays, whose common ancestor with mammals dates from 450 million years ago or earlier.<sup>32</sup>

I'll add one more broad category. In 1948, E. C. Tolman introduced the idea of "cognitive maps," internal representations used by animals in tasks like navigation. Evidence for them was found in behaviors that do not conform to standard behaviorist models of conditioning (especially Skinnerian processes). Examples include shortcut behaviors (which can be seen as spatial versions of "insight"). Map-making in this sense is not restricted to dealing with physical space, and though this is controversial, the construction of some map-like structures is probably not explainable with instrumental learning of the kind discussed here, or other standard associationist algorithms (Gallistel & King, 2010). The discovery of "place cells" and their role in the rat hippocampus has vindicated Tolman's original insights (O'Keefe & Nadel, 1978; Ólafsdóttir, Barry, Saleem, Hassabis, & Spiers, 2015). There's also now evidence for capacities of this kind in some invertebrate groups. The experimental challenge is to rule out explanations of navigation in terms of the use of chemical cues, using a single landmark near the destination, or "path integration." Honeybees, however, can find their way home after several disruptions that would affect these simpler methods (Cheeseman et al., 2014). Some suggestive, though less systematic, work on octopus foraging has been done by Jennifer Mather (1991). Octopuses follow long, looping paths when they forage, and they reliably arrive home even in turbulent waters, often approaching their den from a different direction from the one they went out on. It's not known how this is done, but it might well indicate a capacity for mental mapping of a Tolmanian kind.

When the idea of "mapping" is understood very abstractly, the distinction between the Tolmanian and Skinnerian categories may become blurred in another way. Much work on both classical and instrumental conditioning now looks at "model-based" strategies, in which experience is used not only to associate one event with another but to build a minimal inner model of the dependency relations in the environment that give rise to experienced events.<sup>33</sup> Inner mapping or modeling might thus underly some behaviors usually associated with associationism.

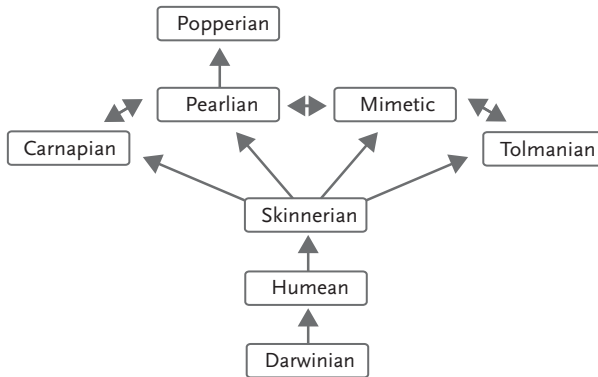
#### 4. Two Trees

More categories might be added, especially in the area of social cognition. A longer discussion might next consider the role of language and its internalization as a

---

32. See also Moore (2004). He counts many independent origins.

33. Doll, Simon, and Daw (2012); Dayan and Berridge (2014).



**FIGURE 2.** A “tower,” dissolving into a network structure with both branchings and rejoinings, making use of the categories discussed in the text. The paths leading out of the “Skinnerian” node are not exclusive; an animal might treat two or more of them at once, and I assume that “lateral” movement between the post-Skinnerian categories is also possible (including movement that bypasses nodes that happen to be adjacent in the figure—those adjacency relationships are not supposed to be significant).

cognitive tool (*Vygotskian creatures*). But I’ll stop the story here, sum up, and consider some morals.

Figure 2 charts the relationships discussed in the last few sections. The main role of the figure is to distinguish between a tower-like structure at the early stages and a branching structure later on. I say “branching,” but the shape is not strictly a tree, because some paths can rejoin after diverging.

There is empirical support for some of the relations represented, though it’s not clear which are merely *de facto* and which are a matter of principle. It’s likely that there are no Skinnerian creatures who are not also Humean and Darwinian. So initially, there is a nesting of categories. At present, though, it appears likely that the categories beyond the Skinnerian level are related in a less cumulative way. It would be hard to show this decisively, because it depends on the absence of capacities in crucial cases. But a possible picture may include relationships like these: bees can map the world, abstract, and show a form of imitation, but they don’t use a rich notion of cause.<sup>34</sup> Octopuses can imitate and map but may have little use for abstraction. Stingrays and tortoises can imitate. Perhaps they can map and abstract, but it would not be surprising if they can’t. I accept, on the other hand, that further work might uncover a single crucial innovation that underlies

34. For bee social learning, see Avarguès-Weber and Chittka (2014).

several of the branches I picture emerging from the Skinnerian category.<sup>35</sup> It's also true that my "Carnapian" category is especially vague; I introduced it partly to make vivid a contrast between mere pattern-recognition and cognition that makes use of a rich concept of cause. The use of this concept of cause is both powerful and uncertain in its relationships to precursor forms.

A further set of questions concern how the structure in Figure 2 relates to the tree of animal evolution represented in Figure 1. Classical conditioning is so widespread that it might have a single origin, or perhaps a couple, in early neural animals. It is also simple in its demands in a way that makes it look relatively easy to invent, however—at least in *some* form—and that weakens the case for a single origin. Instrumental conditioning, in contrast, appears to be more scattered on the tree, though absences of evidence should once again be handled with caution. I've not made much use of molecular evidence in this paper, and this can also bear on questions about single versus multiple origins; Barron et al. (2010) note, for example, that dopamine plays a role in reward-seeking behavior in a great many animal groups. The traits in Figure 2 that go beyond instrumental learning, such as mapping and imitation, also appear to be scattered with respect to the tree. Imitation probably arose independently, for example, in stingrays, tortoises, and perhaps octopuses. Spatial mapping is seen in bees and rats, and has an obvious rationale there, but it may well be absent in many animals who don't need it.

It seems likely that Darwinian evolution built Humean creatures before it built Skinnerian ones, and it built those on a Humean base. Things *might* not be this way—it's possible that they evolved in the other order, or together, and then instrumental learning was lost, diluted, or hidden in a range of animals. But that looks a bit unlikely. Instead, it appears that evolution first built a form of within-generation adaptation that is not a generate-and-test mechanism, but one that works by other means—by "instructive" rather than selective processes, as people used to say.<sup>36</sup> Selection does have a kind of overall primacy in the story—I agree with Dennett about that. But when selection built the first of this sequence of psychological adaptations, it built not more selection but something else. After that, evolution did build another selection process—instrumental learning. Then it built several other things, eventually including the third, internalized, Popperian form of selection. In filling out the story, reward seeking that involves the moment-to-moment shaping of behavior instead of than learning also has

---

35. Heyes, again, argues that merely fine-tuning the input mechanisms can make a Skinnerian creature look quite different (2011).

36. Godfrey-Smith (1996).

to come in somewhere. That is probably older, part of the basis for instrumental learning, and something with its own glimmer of generate-and-test.

Continuing work may also bring more dramatic changes to the picture. I'll mention one example to finish. Brian Dias and Kerry Ressler reported in 2014 that a classically conditioned fear response to an odor in mice was transmitted, through sperm, to children and grandchildren of the conditioned animals. The mechanism appears to involve DNA methylation. Assuming the finding holds up, this is pretty remarkable, and appealingly disruptive with respect to the picture sketched in this paper. Classical conditioning, again, is an "instructive" mechanism, in the terms of the old selective/instructive distinction, and here classical conditioning is part of a means by which a behavior is shaped transgenerationally by a Lamarckian process. (I think it's common to use the term "Lamarckian" too broadly and freely, but this case surely counts.) Two "instructive" mechanisms for change then combine, and a Humean capacity, one step up on the tower, reaches down and affects intergenerational adaptation, otherwise the province of Darwin. Perhaps we are also Escherian creatures, as Dennett's collaborator Douglas Hofstadter has, in a different sense, long argued.<sup>37</sup>

## Acknowledgments

Thanks to Alex Taylor, Bryce Huebner, Tony Dickinson, John Greenwood, and Celia Heyes for help with this paper.

## Works Cited

- Ardiel, E. L., & Rankin, C. H. (2010). An elegant mind: Learning and memory in *Caenorhabditis elegans*. *Learning and Memory*, 17, 191–201.
- Avarguès-Weber, A., Dyer, A. G., Combe, M., & Giurfa, M. (2012). Simultaneous mastering of two abstract concepts by the miniature brain of bees. *Proceedings of the National Academy of Sciences of the United States of America*, 109, 7481–7486.
- Avarguès-Weber, A., d'Amaro, D., Metzler, M., & Dyer, A. (2014). Conceptualization of relative size by honeybees. *Frontiers in Behavioral Neuroscience*, 8. doi:10.3389/fnbeh.2014.00080
- Avarguès-Weber, A., & Chittka, L. (2014). Local enhancement or stimulus enhancement? Bumblebee social learning results in a specific pattern of flower preference. *Animal Behaviour*, 97, 185–191.
- Bain, A. (1859). *Emotions and the will*. London: J. W. Parker and Son.

---

37. Hofstadter (1979).



- Balleine, B. W., Espinet, A., & Gonzalez, F. (2005). Perceptual learning enhances retrospective revaluation of conditioned flavor preferences in rats. *Journal of Experimental Psychology: Animal Behavior Process*, 31, 341–350.
- Barron, A. B., Søvik, E., & Cornish, J. L. (2010). The roles of dopamine and related compounds in reward-seeking behavior across animal phyla. *Frontiers in Behavioral Neuroscience*, 4, 163. doi:10.3389/fnbeh.2010.00163
- Bird, C. D., & Emery, N. J. (2009). Rooks use stones to raise the water level to reach a floating worm. *Current Biology*, 19, 1410–1414. doi:10.1016/j.cub.2009.07.033
- Budd, G. E., & Jensen, S. (2015). The origin of the animals and a “savannah” hypothesis for early bilateral evolution. *Biological Reviews*. Online publication. Retrieved from doi:10.1111/brv.12239
- Campbell, D. T. (1974). Evolutionary epistemology. In P. A. Schlipp (Ed.), *The philosophy of Karl R. Popper* (pp. 412–463). La Salle, IL: Open Court.
- Cheeseman, J. F., Millar, C. D., Greggers, U., Lehmann, K., Pawley, M. D., Gallistel, C., . . . & Menzel, R. (2014). Way-finding in displaced clock-shifted bees proves bees use a cognitive map. *Proceedings of the National Academy of Sciences of the United States of America*, 111, 8949–8954. doi:10.1073/pnas.1408039111
- Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective, and Behavioral Neuroscience*, 14, 473–492. doi:10.3758/s13415-014-0277-8
- Dennett, D. C. (1969). *Content and consciousness*. New York, NY: Routledge.
- Dennett, D. C. (1975). Why the law of effect will not go away. *Journal for the Theory of Social Behaviour*, 5, 169–188.
- Dennett, D. C. (1978). *Brainstorms: Philosophical essays on mind and psychology*. Montgometry, VT: Bradford Books.
- Dennett, D. C. (1995). *Darwin's dangerous idea: Evolution and the meanings of life*. New York, NY: Simon and Schuster.
- Dennett, D. C. (1996). *Kinds of minds: Toward an understanding of consciousness*. New York, NY: Basic Books.
- Dias, B. G., & Ressler, K. J. (2014). Parental olfactory experience influences behavior and neural structure in subsequent generations. *Nature Neuroscience*, 17, 89–96.
- Dickinson, A. (2012). Associative learning and animal cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367, 2733–2742. doi:10.1098/rstb.2012.0220
- Dickinson, A., & Balleine, B. (1994). Motivational control of goal-directed action. *Learning and Behavior*, 22, 1–18. doi:10.3758/BF03199951
- Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, 22, 1075–1081. doi:10.1016/j.conb.2012.08.003
- Fiorito, G., & Scotto, P. (1992). Observational learning in *Octopus vulgaris*. *Science*, 256, 545–547.

- Gagliano, M., Renton, M., Depczynski, M., & Mancuso, S. (2014). Experience teaches plants to learn faster and forget slower in environments where it matters. *Oecologia*, 175, 63–72. doi:10.1007/s00442-013-2873-7
- Gallistel, C. R., & King, A. P. (2010). *Memory and the computational brain: Why cognitive science will transform neuroscience*. Chichester, England: Wiley-Blackwell.
- Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoiding learning. *Psychonomic Science*, 4, 123–124.
- Godfrey-Smith, P. (1996). *Complexity and the Function of Mind in Nature*. Cambridge, UK: Cambridge University Press.
- Gopnik, A., & Schultz, L. (Eds.). (2007). *Causal learning: Psychology, philosophy, and computation*. Oxford, UK: Oxford University Press.
- Greenwood, J. D. (2009). *A conceptual history of psychology*. Boston, MA: McGraw-Hill.
- Haralson, J. V., Groff, C. I., & Haralson, S. J. (1975). Classical conditioning in the sea anemone, *Cribrina xanthogrammica*. *Physiology and Behavior*, 15, 455–460.
- Heyes, C. (2011). What's social about social learning? *Journal of Comparative Psychology*, 126, 193–202.
- Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An eternal golden braid*. New York, NY: Basic Books.
- Hume, D. (1739/2000). *A treatise of human nature*. Oxford, UK: Oxford University Press.
- Kasumovic, M. M., Elias, D. O., Sivalinghem, S., Mason, A. C., & Andrade, M. C. B. (2010). Examination of prior contest experience and the retention of winner and loser effects. *Behavioral Ecology*, 21, 404–409. doi:10.1093/beheco/arp204
- Kim, K. U., Huh, N., Jang, Y., Lee, D., & Jung, M. W. (2015). Effects of fictive reward on rat's choice behavior. *Scientific Reports*, 5 [Article No. 8040]. Retrieved from doi:10.1038/srep08040
- Köhler, W. (1924). *The mentality of apes*. New York, NY: Harcourt Brace.
- Lamarck, J. B. (1809/1914). *Zoological philosophy* (Hugh Elliot, Trans.). London: Macmillan.
- Laurent, V., & Balleine, B. W. (2015). Factual and counterfactual action-outcome mappings control choice between goal-directed actions in rats. *Current Biology*, 25, 1074–1079. doi:10.1016/j.cub.2015.02.044
- Lyon, P. (2015). The cognitive cell: Bacterial behavior reconsidered. *Frontiers in Microbiology*, 6, 264. doi:10.3389/fmicb.2015.00264
- Mather, J. (1991). Navigation by spatial memory and use of visual landmarks in octopuses. *Journal of Comparative Physiology A*, 168, 491–497.
- Moore, B. R. (2004). The evolution of learning. *Biological Reviews*, 79, 301–335.
- Nuttley, W. M., Atkinson-Leadbetter, K. P., & van der Kooy, D. (2002). Serotonin mediates food-odor associative learning in the nematode *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences of the United States of America*, 99, 12449–12454.

- O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford, UK: Oxford University Press.
- Ólafsdóttir, H. F., Barry, C., Saleem, A. B., Hassabis, D., & Spiers, H. J. (2015, June). Hippocampal place cells construct reward-related sequences through unexplored space. *ELife*, 4, e06063. Retrieved from <http://dx.doi.org/10.7554/eLife.06063>
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, England: Cambridge University Press.
- Perry C., Barron A., & Cheng, K. (2013). Invertebrate learning and cognition: Relating phenomena to neural substrate. *WIREs Cognitive Science*. doi:10.1002/wcs.1248
- Peterson, K., Cotton, J., Gehling, J. G., & Pisani, D. (2008). The Edicaran emergence of bilaterians: Congruence between the genetic and the geological fossil records. *Philosophical Transactions of the Royal Society B*, 363, 1435–1443.
- Rescorla, R. A. (1967). Pavlovian conditioning and its proper control procedures. *Psychological Review*, 74, 71–80.
- Rogers, A. R. (1989). Does biology constrain culture? *American Anthropologist*, 90, 819–831.
- Rudin F. S., & Briffa, M. (2012). Is boldness a resource-holding potential trait? Fighting prowess and changes in startle response in the sea anemone, *Actinia equine*. *Proceedings of the Royal Society B*, 279, 1904–1910.
- Ryan, J. F., Pang, K., Schnitzler, C. E., Nguyen, A., Moreland, R. T., Simmons, D. K., . . . & Baxeavanis, A. D. (2013). The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. *Science*, 342, 124259. doi:10.1126/science.1242592
- Shanks, A. (2002). Previous agonistic experience determines both foraging behavior and territoriality in the limpet *Lottia gigantea* (Sowerby). *Behavioral Ecology*, 13, 467–471. doi:10.1093/beheco/13.4.467
- Shettleworth, S. J. (2012). Do animals have insight, and what is insight anyway? *Canadian Journal of Experimental Psychology*, 66, 217–226. doi:10.1037/a0030674
- Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search*. Cambridge, MA: MIT Press.
- Taylor, A. H., Cheke, L. G., Waismeyer, A., Meltzoff, A. N., Miller, R., Gopnik, A., . . . & Gray, R. D. (2014). Of babies and birds: Complex tool behaviours are not sufficient for the evolution of the ability to create a novel causal intervention. *Proceedings of the Royal Society B*, 281, 20140837. doi:10.1098/rspb.2014.0837
- Terao, K., Matsumoto, Y., & Mizunami, M. (2015). Critical evidence for the prediction error theory in associative learning. *Scientific Reports*, 5 [Article No. 8929]. Retrieved from doi:10.1038/srep08929
- Thonhauser, K. E., Gutnick, T., Byrne, R. A., Kral, K., Burghardt, G. M., & Kuba, M. J. (2013). Social learning in cartilaginous fish (stingrays *Potamotrygon falkneri*). *Animal Cognition*, 16, 927–932.

- Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press.
- Tomasello, M., & Call, J. (1997). *Primate Cognition*. Oxford: Oxford University Press.
- Wells, M. J. (1968). Sensitization and the evolution of associative learning. *Symposium on the Neurobiology of Invertebrates 1967*, 391–411.
- Wilkinson, A., Kuenstner, K., Mueller, J., & Huber, L. (2010). Social learning in a non-social reptile (*Geochelone carbonaria*). *Biology Letters*, 6(5): 614–616. doi:10.1098/rsbl.2010.0092
- Woodward, J. (2005). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

## 8.2 REFLECTIONS ON PETER GODFREY-SMITH

Daniel C. Dennett

Peter Godfrey-Smith recognizes the big claim behind my Tower of Generate and Test: “There is a single pattern essential to all processes of adaptation, cognitive improvement and R & D in the broad sense of the term” (p. 225). Wherever there is design (in the broad sense that includes all the design in the biosphere), it is the result of design improvement achieved by processes that echo the fundamental Darwinian process in which candidates are generated, tested, and culled without the aid of skyhooks. Even works of inspired human genius turn out on examination to owe all their excellence to myriads of mindless, mechanical processes that were themselves honed by trial and error, and invoke trial and error, at different tempos and with candidates of hugely varying complexity. Design improvement is always a matter of exploiting encountered information to ratchet up the powers of the design. This has been my preoccupation since 1969, as Godfrey-Smith notes, and *BBB* serves up quite a few embellishments and improvements to my earlier presentations of these ideas, but what he shows is that I missed some golden opportunities along the way. I welcome the many further insights he has winkled out of my earlier versions.

First, he is right to be leery of my building a Tower where any good Darwinian should plant a Tree, but my reflections on how I could have made this rookie mistake led me to a fairly satisfying rationale, part of which Godfrey-Smith articulates himself. Sometimes in nature there *is* a “nesting” of mechanisms, at least for a period, and perhaps my Tower is such a structure. I think it is. (By the way, I agree that I haven’t been as clear on this as I should, but I meant that Skinnerian creatures were *merely* Skinnerian creatures, while Gregorian creatures such as us (and only us) are also Popperian, Skinnerian and, of course, Darwinian.) But another part of the rationale, I have decided, is that the Tower

represents major differences in power, not necessarily “major transitions in evolution.” Think of the Chomsky hierarchy of ever more powerful computers able to accept ever more complex formal languages. The Chomsky hierarchy is a Tower, not a Tree. There is no clear sense in which one type of computer (e.g., a nondeterministic pushdown automaton) is a descendant of any other type.

One of the most interesting fruits of Godfrey-Smith’s discussion is the illumination of the curious status of classical Pavlovian (Humean—in Godfrey-Smith’s terms) conditioning. Since it doesn’t depend on reward it at first seems like a dubious gift of nature. In what way, for instance, is Pavlov’s dog benefited by acquiring its bizarre salivation habit triggered by bell ringing? When Godfrey-Smith tries to imagine creatures who are Pavlovian but not Skinnerian, he arrives at a provocative suggestion:

An animal could be completely blind to reward—completely unable to distinguish between good and bad outcomes—and yet be able to learn by classical conditioning: it could still learn that when *X* happens, *Y* happens soon after, and it could come to produce its *Y*-adapted behavior in response to *X* as well. It might not register the good consequences of that production of a *Y*-adapted behavior; only blind Darwinian processes might assign the appropriate behavior to stimulus *Y*. (p. 233)

This seems to be a recipe for extending the class of stimuli that provoke an “instinctual” response beyond the class that was heretofore endorsed by natural selection. What reason is there to think that this relaxation of the boundaries would usually be beneficial? Recalling the Frenchman being interviewed in the joke, who says that cars remind him of sex, while skyscrapers remind him of sex and that, yes, food also reminds him of sex—“*everything* reminds me of sex!”—suggests that a plethora of associations would not always be a cognitive enhancement. And yet Hume’s associationism does seem to be a step in the right direction, a basic mechanism one would want to have in one’s kit. How do we reconcile this? I suggest that what we need to add to the system gifted with Humean association is precisely the ability to distinguish between good and bad outcomes, but not necessarily by the mechanism of Skinnerian reinforcement in the hard knocks of the external world. Hume, after all, was thinking about human thinking, about the crowds of ideas (and “impressions”) milling around in people’s minds, and if we see such a mind as an arena of competition, with “good” combinations eventually dominating “bad” combinations, then classical conditioning would serve as a Generator of Diversity, promoting into competition a steady stream of hopeful associations, that were then subjected to a Darwinian process of selection, culling

out the spurious. The refinement process might be hugely inefficient, promoting a lot of attractive nonsense along with the good stuff, but the very fact that an association had been formed would testify that there was a pattern, a better-than-chance relationship between two ideas (or between unconditioned stimulus (US) and conditioned stimulus (CS), in Pavlovian terms). We might recast all this usefully into the Bayesian perspective—see Clark (chapter 7.1) and my Reflections (chapter 7.2), this volume.

When he turns to Popperian creatures, Godfrey-Smith surmises that “Dennett mishandles his own hierarchy.” Yes. He’s right, and this had been in the back of my mind for some time as a weak spot that needed attention. I agree that I should have set the bar high, so that “an internal consequence-assessor” is the innovation that matters, but having said that, I think that we need to recognize that, as usual, there will be all manner of intermediary forms, *sorta* consequence-assessors in some regards and not in others.

What of the population of further creatures: Tolmanian, Pearlian, Carnapian, Vygotskian, Escherian? Each of these vividly captures a different way of achieving cognitive competence, and what is particularly useful is seeing the possibility that, for instance, the first three might well be found among nonhuman, non-language-using species. I’m not sure how fruitful the taxonomy would prove, but it might be a winner (as long as we banish the Köhlerian creatures—I agree with Godfrey-Smith that this not a useful variety). But I want to deliver a preemptive strike against a theme that is *not expressed* by Godfrey-Smith, but that some over-eager readers might see as supported by his discussion:

There are many different ways of being intelligent, and the understanding of the chimpanzee, the elephant, the electric eel, the corvid . . . is not *less* but just *different*. We humans shouldn’t think of ourselves as cognitively superior to other creatures.

This celebration of difference is all very well, but the glaring fact remains: no sooner do we human beings, we Gregorian creatures, uncover another Good Trick in some other species than we adopt it by designing prosthetic perceptual apparatus: bat-like sonar and radar, bloodhound-like chemical olfaction systems, and the rest. With our thinking tools, we can not only mimic the prowess of other species; we can surpass it, often by orders of magnitude. That is just perception, of course. What about tactical thinking, and wisdom? As a marine biologist once said, “Dolphins may be brilliant, but if so, they are doing a brilliant job of concealing that fact from us.” Gregorian creatures, by pooling the design experience of the whole species, have cognitive powers that can take any newfound

cognitive powers in some of our fellow humans or in other species and arrange to enjoy their benefits ourselves. In this regard, Gregorian creatures are a little like Universal Turing machines—their competences are practically limitless.

### Work Cited

Dennett, D. C. (2017). *From bacteria to Bach and back*. New York, NY: W. W. Norton.



# 9.1

## MOTHER CULTURE, MEET MOTHER NATURE

Luc Faucher and Pierre Poirier

A scholar is just a library's way of making another library.

—DENNETT, "Memes and the Exploitation of Imagination"

Here is a fairy tale of sorts. Once upon a time, a very long time ago, a molecule found a way to replicate. Its descendants multiplied and populated the Earth. Because of their venerable age, let's call these replicating molecules "the old replicators." These old replicators eventually found that they could sometimes out-replicate their rivals by building *survival machines* to interact with the world for them. By providing these survival machines with control centers, they made the process of replication highly efficient. And this eventually fostered the creation of a new type of replicator, made out of patterns within these control centers. These "the new replicators" eventually gave the survival machines the ability to understand how the world works. This was a very good thing for it gave "new-replicator-infected" machines tremendous control over other machines. As a result, they multiplied and came to populate every corner of the Earth.

One of these machines, call it Mendel, discovered how the old replicators worked. And as a result, many of the new replicators that populated Mendel's control center came to infest other machines. Eventually, these new replicators mutated, leading to a machine called Dawkins, which was thoroughly infected by Mendel's new replicators, to work out how the new replicators worked, as well. The Dawkins machine's new replicators infected many more machines. But they might have been wiped out, if it had it not been for a machine called Dennett, which reinvigorated them by creating robust and aggressive copies, and spreading them more efficiently.

But every good story needs a turn of fortune. And as these new replicators infested more machines, some of those infected machines

came to view the working of the old replicators slightly differently than Mendel had. Mendel's new replicators, they felt, only told part of the story. A new way to view the old replicators was now needed; and the machines worked hard to produce that revised story, hoping that it, too, would replicate and become the new way to view the old replicators. But if the view that is now beginning to replicate is right, we may have to change our view of how new replicators work, as well. That is what this paper is about.

According to Daniel Dennett, Culture is akin to (biological) Nature on two related fronts. First, Culture is the result of a Darwinian selection process acting on physical structures that can be ascribed informational content by those seeking to understand the process and its outcome. Second, both Culture and Nature can be seen as intentional agents ("Mother Nature" and "Mother Culture") who act on the basis of practical inferences to further their goal of adapting organisms to their environment. For Dennett, cultural meaning is always ascribed from the Intentional Stance. For the Darwinian algorithm to generate cultural structures and processes that can be interpreted through the Intentional Stance, its elements must be interpretable structures, as genes are thought to be (Dennett, 1995).

Recently, our conception of genes and our knowledge of the types of Darwinian algorithms one finds in Nature have been transformed under the impulse of new fields like "genomics" but also through the careful study of biological phenomena such as immunity and disease. Any science that aims to study culture with the tools of biology, as memetics does, cannot ignore the changes that are currently stirring biology. Those transformations may lead memetics to drop (or to profoundly modify) some of its claims that are now no longer seen as the best descriptions of cultural processes, because richer models from biology provide better ways to understand the processes involved; but, they may also help further other claims made by memetics. It is, we claim, only by staying open to developments in biology that memetics has a chance to eventually reach a scientific status.

We first draw a cautionary lesson from what research on the immune system tells us about Darwinian algorithms (section 2). Second, we turn to the concept of "gene," or what's left of it, in the context of postgenomics to see how adopting a similar conception for "meme" might help memetics to answer some objections that have been formulated against it (section 3). To make our point, we will consider one such objection (that of Maurice Bloch) and provide an example of the explanatory gains that memetics could achieve if it were to adopt a conception of a meme along the lines of the new conception of gene. But first, we want to state what we believe Memetics to be at the present time and to review

some of Dennett's motivations for adopting it as a way of thinking about culture (section 1).

## 1. Old-Fashioned Memes (or, the Old New Replicators)

### 1.1 *The Virtue of Being Pluralist*

In the late 18th and early 19th centuries, the idea that species transform into other species (a doctrine called *transformism*) emerged in opposition to the purportedly Aristotelian view that species are fixed (a doctrine called *fixism*). Darwin's theory of evolution by natural selection aimed to show that transformism is true by explaining how the change from species to species occurs (Depew & Weber, 1995). Although it is aimed at explaining species change, Darwin's theory explains by extension how a given species came to have the traits it currently has. In the process, it has also helped redefine what it is to be a species (Sober, 1980).

There is no debate among anthropologists and cultural theorists over whether cultures are fixed or change: one simply has to look at the rapid change occurring in America regarding LGBTQ rights to see *live* cultural transformation (although the rate of change may be faster in some cultures or time periods). Memetics, as we see it, is any attempt at explaining cultural change as a type of selection process. Thus defined, memetics is not a theory or an explanation, but a conjecture about the type of mechanism that might explain cultural change. Since we will refer later on to stricter definitions of the term, that is, actual proposed mechanisms (or "mechanism sketches"; Craver, 2007) to explain cultural change, let's call this conjecture "Memetics-C," for "Memetics-as-a-Conjecture." This paper is not addressed to those who oppose Memetics-C, but to those who, like us, accept it but believe that the time has long passed for Memetics-C to give rise to *actual* explanations of specific cultural changes. To use Craver's (2007) helpful vocabulary, there are many *how-possibly* explanations of cultural change, but precious few *how-actually* explanations (but see Hull, 1988, for a well-developed *how-actually* explanation). The fact that it has proven so difficult to go beyond *how-possibly* explanations is, we believe, one reason why Memetics has such bad press nowadays. We hope here to help change this situation.

Perhaps a good place to begin is by looking at the type of mechanism conjectured by Memetics-C to be relevant to the explanation of cultural change: a process of selection that optimizes some quantity. Some view Memetics-C's main conjecture as a kind of analogy. Just as Rutherford (1911) explained the structure of the atom by analogy to that of the solar system, so would memeticists explain cultural change by analogy to Darwin's explanation of species change. However, Dennett, like most prominent advocates of Memetics-C, argues that the selection mechanism responsible for cultural change is not merely analogous to the

selection mechanism responsible for species change: both biological change and cultural change *are the result of the same type of universal process*, a view sometimes called Universal Darwinism (UD):

Meme evolution is not just analogous to biological or genetic evolution. It is not just a process that can be metaphorically described in these evolutionary idioms, but a phenomenon that obeys the laws of natural selection exactly. (Dennett, 1990, p. 128)

According to UD, natural selection occurs whenever the material conditions for a certain kind of selection process are present.<sup>1</sup> Thus understood, this selection process is not especially biological: it is a substrate-independent *Evolutionary Algorithm* (EA), whose first discovered instantiation just happened to be biological (Campbell, 1974; Dawkins, 1976; Dennett, 1995; Hull, 1988). As Dennett (1995) emphasizes, EAs form a class of algorithms that share some general features but whose details, we might add, may be quite different. Take *greedy algorithms* as an example. Greedy algorithms are optimization algorithms that look for the largest short-term gain without regard for long-term considerations. They are therefore quick to find a solution, but this does not guarantee it will be the optimal solution. Beyond this very general description, the details of greedy algorithms can vary wildly (since “looking for largest short-term gain” and “disregarding long-term considerations” can both be implemented in a variety of ways). Similarly, EAs as a class include algorithms that may vary a lot when one looks at the precise detail of their logical structure (although they will share those important features in virtue of which they belong to the class).

Another feature of EAs, rightly emphasized by Dennett, is their medium independence. Any formal system can be instantiated through (humanly manipulated) pen and paper, copper wires and vacuum tubes relays, silicon chips, or any other physical medium that respects the formal system’s identity conditions (its set of tokens, starting state, and transformation rule(s); see Haugeland, 1985). Likewise, an EA can be implemented in any medium that can be fashioned to respect its identity conditions, no matter *how* the medium in question has been fashioned: by the laws of physics and chemistry (genetic evolution), by the laws of biochemistry (immune-system evolution), by the laws of neuroscience (neuronal Darwinism), by the laws of neuroscience and psychology (cultural

---

1. Along with other types of change—notably, changes that occur in the immune system (Hull et al., 2001) or in the brain (Edelman, 1987; Fernando et al., 2012). Some (e.g., Campbell, 1974), wishing to reserve the terms “Darwinism” and even “evolution” for the type of change Darwin’s theory was meant to explain, call the general class “selection theories (of change).”

evolution), or by humans programming computers (genetic algorithms and genetic programming):

The power [of an algorithm] is due to its logical structure, not the causal power of the materials used in the instantiation, just so long as those causal powers permit the prescribed steps to be followed exactly. (Dennett, 1995, p. 51)

It is, however, easy to get carried away by the medium independence of algorithms. (Years of excessive claims in the philosophy of mind referring to “multiple realizability” can attest to that!) The “materials used for instantiation” do matter epistemically when one’s aim is discovering which algorithm is instantiated by a given mechanism. Since EAs are a class of algorithms, knowing that the mechanism responsible for cultural change instantiates an EA says both a lot and little about the actual mechanism itself (just like knowing that a piece of code instantiates a greedy algorithm says both a lot and little about that piece of code). In modern computers, the algorithm can be decided beforehand (of course, testing and experimentation will be needed to determine whether the instantiated algorithm actually performs what we wanted it to perform or needs adjustments). With memetics, the algorithm is unknown beforehand, and the task of scientists is to discover which algorithm is actually instantiated in a mechanism.

Given the forgoing, there are two ways to proceed in order to turn sympathy toward Memetics-C into *how-actually* memetic explanations of cultural change: the greedy way and Darwin’s way. Here is the greedy way (an actual instance of a greedy algorithm): (1) find an instance of a well-worked-out EA; (2) declare that the mechanism responsible for all cultural change instantiates *that* EA; (3) choose one cultural change and proceed to explain it as the result of an EA mechanism; (4) repeat. Here is Darwin’s way: (1) choose one cultural change, knowing that an EA is responsible for some aspects of that change; and (2) study that change to uncover the exact details of the selection process or processes that brought it about. (By calling this method “Darwin’s way,” we do not mean to claim that Darwin had any clear idea that EAs were a class of algorithms but simply that he furiously studied any change he sought to explain.)

As we see it, early advocates of Memetics-C got too greedy too soon—as the greedy phase of an optimization process is often best introduced late and progressively. Given that EAs form a class of algorithms that can be realized differently in different situations, advocates of Memetics-C should not prejudge the nature and number of specific EAs necessary to explain cultural change. Culture is a complex phenomenon, with no assurance of ontological or conceptual unity. Accordingly, some cultural changes may be best explained by one specific EA; whereas others

are best explained by another. Since the identity conditions of formal systems include their set of tokens, that this, the set of “objects” over which the system’s transition rule is defined, the point just made naturally extends to the objects over which EAs operate. These must share some features (to cite famous examples: fidelity, fecundity, and longevity, but also digitality, location, etc.) for the EAs transition rule to apply (the rules of Classical Logic do not apply to bananas; the rules of EAs generally do not apply to dust particles). In any domain where the objects over which the EAs rules are defined may differ importantly (and remember: whether they do or not should not be prejudged), one might find various EAs working side by side, since each set of objects, given its features, may have favored the use, and indeed evolution, of a different EA (as some operations are impossible or much more difficult on certain types of objects). For instance, the fact multiplication with Roman numerals is much more difficult than with Arabic numerals is responsible for the use, indeed (cultural) evolution, of multiplication algorithms defined for Arabic numerals). As we see it, advocates of Memetics-C have also been too greedy too soon in settling too soon on a specific conception of a meme.

We believe that many of the thorny problems that have plagued the field of Memetics result from the ill-advised narrowing of the scope of EAs (including their objects) and that, accordingly, adoption of a more pluralistic approach in pursuing research motivated by Memetics-as-a-Conjecture (i.e., Memetics-C) helps address many of these problems. But before we address them specifically, it will help to acknowledge Dennett’s reasons for defending his particular way of pursuing such research, which we call “Memetics-D” (for Dawkins, or Dennett<sup>2</sup>). Memetics-D is characterized by a particular (narrow) way to understand both the *processes of selection* and the *replicators* on which they are acting. In this paper, we want to argue that Memetics-D should give way to a Memetics-P (for “pluralistic”).

## 1.2 The Expected Benefits of Adopting a Memetic Perspective on Culture

Dennett’s first motivation for adopting Memetics-D derives from the revolutionary implications of Darwinian ideas. As Dennett (2006b, p. 119) puts it: “Darwin’s dangerous idea [the theory of evolution by natural selection] amounts to nothing less than a reframing of our fundamental vision of ourselves

---

2. Although Dennett (1995) has its pluralistic moments, the main focus is on the Dawkinsian, selfish meme and so we will take him here to be a defender of Memetics-D (but see Dennett, 2011, for signs of growing pluralism).

and our place in the universe.” Accepting Darwinian ideas in biology has turned our world upside down. We can’t see ourselves as on top of Nature anymore, but as just one very eccentric branch of the tree of life (Gould, 1996). By expanding the reach of Darwin’s idea to culture, we can hope for a similar revolution in our self-conception. We like to think that we are the creators of our ideas, in charge of weeding out the bad ideas, of identifying the good ideas and sharing them with others, etc. In accepting Memetics-D, we can see how to give up on the idea that we are the person in charge, that we are authoritative on what seems to be the closest to us, that is, the content of our mental lives.

Memetics has another impact on the way we should think of ourselves. As the bacteria in our digestive tract, without which we could not digest food, are considered as part of “us” (Hutter et al., 2015), memes are not the “other”; they are us (Dennett, 1990, p. 133). They are what motivate us to act, they affect the way we perceive, judge or think about some objects or some people. Memes have even created this new type of biological entity that we call a “person.” Being a person, this is at least what some philosophers have argued (for instance, Frankfurt [1971] or Watson [1975]) is being able to identify with a subset of the memes that populates our mind (i.e., being a person consists in wanting or desiring wholeheartedly some of our first-order wants or desires). The person we are, according to this picture, depends on which memes are put in the “driver’s seat”<sup>3</sup> (Dennett, 1995, 367); the idea that it is important to identify with some memes rather than others is also, according to Dennett (2001, p. 322), a meme (he calls it a “meme-borne attitude”). At the end, we, as persons, are memes all the way down. There is no (real, meme-independent) self inside us, no ghost in the machine. And there goes another myth down the drain, courtesy of Darwin (and Dennett).

The second benefit coming from the adoption of Memetics, according to Dennett, is the potential of accrued consilience between the biological and the social sciences. Biology and social sciences had a tumultuous relationship: from a rather peaceful interpenetration of the domains in the Nineteenth century (as exemplified by Edward Tylor’s anthropological theory), it went through a process of autonomization of the social sciences to a violent clash in the 70s in response to sociobiological imperialistic reductionism. Though the dust has settled quite

---

3. Memes that got in the driver’s seat are the one with which we identify. There are many memes in ones mind (for instance, the memes of “being fit,” of “taking care of one’s appearance,” etc.) that just “hang out” there, so to speak. They are wishes. They never are acted upon. We never really identify with them. Other memes get to “control” our life, they set our priorities: these are the ones with which we identify and they define the kind of person we are (Dennett, 1995, p. 168). Now, and to be clear, some of the “driving” is perform by mental structures that are not in the driver’s seat (they might as well be seen as defining who we “really” are; see Faucher, 2016). This will be the topic of section 3.3.



a bit since then, proposals have multiplied that tried to re-connect biology with social sciences (for a description of the major proposals, see Laland & Brown, 2002). For people like Dennett, “any theory of culture change worth more than a moment’s of consideration will have to be Darwinian in the minimal sense of being *consistent with* the theory of evolution by natural selection of *Homo sapiens*” (2000, p. ix). In a sense, this is a very minimal requirement: being consistent requires not contradicting the tenets of Darwinism, but it does not require including them in the theory. A theory respecting this constraint could, for instance, not mention or even not include at all Darwinian principles in its core. Memetics-D does something more: it includes those very principles in its theorizing of cultural changes. Memetics-D is not the only theory to try to reconnect biology with social sciences: for instance, in recent years, evolutionary psychology and gene-culture co-evolutionary theory had tried to do so as well (Tooby & Cosmides, 1992; Richerson & Boyd, 2005; Henrich, 2015). But Memetics-D does present, according to Dennett, an idea that is overlooked by the other models or theories: that is, the idea that cultural items might not be good at anything besides replicating themselves. There might (sometimes) be no explanation for their adoption or replication in terms of the survival advantages it would give to their vehicles (i.e., humans). This is not to say that a cultural item (an innovation) is never beneficial for its host (or for both the host and the item at the same time); it is only to say that we should not presume this will always be the case.

This view of cultural evolution not only completes other evolutionary theories of culture in its pursuit of greater consilience between the sciences, it also modifies the dominant model of cultural evolution in social sciences. According to this model (which Bloch calls the “Hamlet model,” following Dreyfus & Dreyfus [1986, p. 28]), “culture is composed of various valuable practices and artifacts, inherited treasures, in effect that are recognized as such (for the most part) and transmitted deliberately (and for good reasons) from generation to generation” (Dennett, 2009, p. 3). On this view, the rationality of agents would be in the forefront of the explanation of cultural evolution (understood as cultural changes). Adopting the meme’s perspective does not lead to a complete overthrow of this picture. Rational choice by agents explains some patterns of change, but we should not presume that it explains them all.

The final (and related) reason to adopt a memetics perspective is epistemological. Memetics is seen as a corrective to a faulty epistemological view according to which our epistemological behaviors are primarily guided by our norms: for instance, that our reason for adopting a particular idea is because it is true (or likely to be true). But how are we to explain the numerous deviations from this idea documented by psychologists? This is where Memetics-D comes in: “The meme’s-eye view purports to be a general alternative perspective from which



these deviations can be explained” (Dennett, 1990, p. 130). From the memetics perspective, the reproductive success of a meme is independent of its “epistemological virtue,” defined either as truth, scientific or aesthetic excellence. Dennett (1990, p. 130) further argues that “only if meme theory permits us better to understand the deviations from the normal scheme will it have any warrant for being accepted.”

We will consider these benefits as desiderata in the evaluation of the changes we propose to Memetics-D. In other words, we argue that, to be acceptable by Dennett (or any other proponents of Memetics-D), Memetics-P should allow the theoretician of culture to reap the same expected benefits as Memetics-D.

## 2. New Kids on the Block: Selection Algorithms Galore

We take the moral of Section 1 to be that there are arguably benefits to adopting Memetics-C in the study of culture. Unfortunately, though perhaps understandably, early proponents of the conjecture were too greedy in choosing one specific selection algorithm as best suited to capture interesting regularities and foster a better understanding of cultural phenomena (instead of taking Darwin’s route of furiously studying the process that lead to change). It would be foolish of us to propose a replacement algorithm for the one greedily chosen by early adopters of Memetics-C. Our objective, rather, is to show the value of taking Darwin’s route and keeping the set of selection algorithms open when attempting to cash out a selectionist conjecture to explain a phenomenon. To this end, we will rehearse the well-known case of the immune system (Hull, Langman, & Glenn., 2001) to show how multiple selection processes can contribute to changing the state of a system.

To adequately fight off infections, a healthy immune system must contain about a million different antibodies. Each antibody is unique and functions to identify pathogens, marking them out for destruction by binding to them,<sup>4</sup> providing a target for the organism’s “defenses” (specifically, cells and enzymes capable of killing the pathogen) to attack the marked pathogen. In its war against pathogens, the immune system must solve three problems: (a) identifying many types of existing pathogens (prions, viruses, molds, and bacteria), and (b) identifying new types of pathogens that evolve. Brute force solutions, where species would keep a record for the blueprint of every necessary antibody in their genome, are unavailable (in a reasonably sized genome). And the strategy of evolving a new

---

4. The part of the pathogen to which an antibody binds to is called an *antigen*. Not all of the antibody binds however to the pathogen: the part (of the antibody) that does is called the *paratope*, and the part that doesn’t the *epitope*.

antibody every time a new pathogen appears is also unavailable, as the generation of pathogens is more rapid than the evolution of its host, so pathogens could wipe out a species before the antibody could evolve. Finally, (3) antibodies should mark all and only pathogens. Although marking can never be perfect, an immune system that “forgot” to mark pathogens (or “forgot” marked pathogens) would be susceptible to more infections and pathogen-borne diseases, and an immune system that marked nonpathogens would eliminate some of the organism’s own cells (autoimmune diseases) as well as other potentially useful cells (recall that “the organism” is made up of about as many “alien” cells—bacteria and other commensals—as it is of descendants of the original cell formed by the two gametes). To solve these problems evolution has homed in on a mix of selection processes all contributing in their own way to the immune system’s function.

In mammals, newborns receive their immune system through two different paths, involving different types of selection processes.<sup>5</sup> Along a replication/selection process familiar to modern evolutionary biology, newborn mammals receive from each of their parents replicated copies of genes selected because they coded for proteins implicated in the construction of successful antibodies. As is now well known, newborns also receive immune factors (antibodies) from their mother’s milk, a process known as “passive immunity.” Unlike their parent’s genes, newborns do not receive *copies* of their mothers’ antibodies but a subset of their mothers’ antibodies. The process is based on reproduction (cell division) and transmission rather than replication. At this point in the construction of the infant’s immune system, replication is only involved for one of its sources and so explaining the current state of the immune system will have to refer to two different selection processes. This, strictly speaking, is sufficient to make our point concerning the value of adopting Darwin’s way over the greedy way. But things only get more complicated from here.

The first noteworthy addition is that the passive immunity is not so passive after all. It is now believed that newborns actively shape their (not-so-)passive (anymore) immunity (Breakey, Hinde, Vallengia, Sinofsky, & Ellison, 2015). Some researchers now believe that illness in breastfeeding infants may increase the production immune factors (notably milk lactoferrin) in their mothers’ mammary glands, which are then transmitted to the infant via breast milk. Little is known about the whole process at present, but by introducing a loop through

---

5. Fetuses have a different immune system from the newborn’s (and a fortiori the adult’s), deriving from a different stem cell line from that of newborns and adults. It is thought that the main function of this other immune system is to “teach” the foetus to be tolerant of its own cells and that of its mother (Mold, Venkatasubrahmanyam, Burt, Michaëlsson, Rivera1, Galkin, et al., 2010). We will concentrate here on the newborn and the adult systems.

the mother (in which the infant gets its mother to produce immune factors to fight its disease), it adds to the complexity of the selection algorithm that explains the changes and state of the infant's immune system.

Antibodies are produced in the soma of a type of cell called "B-cells." It is in those cells that genes are expressed to construct antibodies. Each antibody is made up of two parts: an L-chain and an H-chain (where L and H, respectively, stand for "Light" and "Heavy"), each contributing equally to the specificity of the antibody. Every chain (whether L or H) contains a variable region V and a constant region C. The V-region is by definition the region that contributes to the antibody's specificity. About 100 gene segments code for the V-region of the L-Chain, and the same for the V-region of the H-chain (by contrast, only one gene segment each codes for the C-region of the L and H chains). This means that these roughly 200 gene segments can code for 10,000 distinct antibodies. To make sure that joining errors are minimized, each B-cell is made "functionally haploid" (Hull et al., 2001) so that each B-cell contains only one specific L-H combination (out of the possible 10,000). When the concentration of a given pathogen exceeds a certain threshold (meaning that the pathogen has started replicating within the organism), the 200 or so gene parts that code for parts of the antibody molecule (mainly the V-part of their L and H chains) are activated, and the parts thereby produced are properly joined. The bloodstream is thus flooded with B-cells of 10,000 different types (one antibody specificity per B-cell).

Variation, the first important moment of the selection process, is thus produced. But unlike genomic evolution, variation at this stage of the immune response is produced in a burst, a "Big Bang" to use Edelman's term.<sup>6</sup> As a selection algorithm, this is different from an algorithm that slowly produces variation (the two algorithms will not explore the space of possible solutions in the same way, leading to a host of different algorithmic properties: differences in how much variation and how variation is temporally distributed). Note, however, that there is controversy among partisans of evolutionary models of immunology. Edelman believes in a true Big Bang model where all the variation needed is produced at once; whereas Hull et al. present a replication-interaction cycle model, where the burst is produced iteratively throughout the cycle. Once again, these are different algorithms. Hull et al.'s cycle model offers more space for controlling the unfolding of sets of antibodies than does a strictly Big Bang model. Nevertheless, selection algorithms can well accommodate one-time creation events, even if it turns out that this is not how genomic evolution unfolded. Moreover, all types

---

6. Stephen Jay Gould has argued that biological (genomic) evolution often comes in burst following, for example, mass extinction phenomena. But no one has proposed that anything close to a true Big Bang (tens of thousands of simultaneous speciation events) occurs at that level.

of intermediaries between one-time creation models and the more continuous replication-interaction cycle models are possible (for instance models where aggressive and laid-back variation-creation phases alternate).

Antibodies are produced in B-cells, which like most somatic cells undergo constant reproduction in the organism. However, somatic cells do not reproduce by replication but by division (Godfrey-Smith, 2011). Although these will be involved in a selection process (indeed many), it will not be one based on replication. Three processes occur while B-cells reproduce. First, not all B-cells reproduce at the same rate. B-cells that react to the organism's own cells are selected against (the mechanism by which they are selected against is not well known; Hull et al., 2001). That (yet to be fully worked-out) mechanism neutrally selects B-cells that do not react to the organism's own cells. Second, B-cells that react to pathogens are strongly selected for. Those B-cells that respond to the pathogen, that is, those whose antibodies bind to the (or one of the) pathogen's antigens, divide and thus reproduce, thereby secreting in the bloodstream large amounts of antibodies that can bind to pathogens. As a result, the organism's defenses attack the pathogen, which, if the immune response is sufficient (it is unfortunately not always sufficient—take the Ebola epidemic of 2014 in West Africa), will rid the organism of the pathogen.

As a result of these two mechanisms, a three-pronged selection process is in place, leading to differential reproduction of B-cells according to the target of their specificity: B-cells that bind to the organism's own cells are selected against, B-cells that react to pathogens are strongly selected for and B-cells that react to neither the organism's cells or to pathogens are neutrally selected. After a while, the bloodstream has a high concentration of pathogen-binding antibodies, a medium concentration of neutral B-cells and a low (or nil) concentration of anti-self B-cells. However, that is not the end of the organism's immune response to pathogens. While they reproduce, those B-cells that are selected for undergo mutations at a high rate (6 orders of magnitude faster than the normal mutation rate). As a consequence, variation is once again<sup>7</sup> introduced in the system: many slightly different V and C parts of L and H chains are generated, each leading, when fusion occurs, to a population of slightly different antibodies, capable of perhaps better binding to the pathogens antigen(s) or of tracking in real time the possible mutations in the pathogens (that is, capable of tracking the real time, on-line evolution of the pathogen). This is a case where the somatic evolution of a slow reproducing

---

7. For those who keep track, this is the third embedding of variation generation: phenotypic, at the initial onset of the immune response and now to fine-tune the response and track in real time the potential on-line evolution of the pathogen.

organism (or one that does not reproduce as fast as its pathogens) matches the genetic evolution of a rapidly evolving organism. The timing of the two evolutionary processes is important for the slower evolving organism to be able to fight infection by a fast-evolving organism, and natural selection has found a way to match them. In introducing variation, however, it is possible that B-cells create new antiself antibodies, but the B-cells that carry them will be selected against by the (yet to be fully understood) mechanism that eliminates antiself antibodies.

Let's call the B-cells that result from the activation of the 200 gene components already present in the organism prior to infection the "prior B-cells" and those that result from the mutation of selected-for B-cells the "post B-cells." Prior to the current infection, the distribution of B-cell specificity in the bloodstream reflects the organism's history of encounters with pathogens in addition to the antibodies it received from its mother. This is the pre-infection B-cell population. Note the complex cascade of events triggered by pathogens. When the concentration of pathogens reaches a certain level, the pathogens (1) cause an increase in B-cell reproduction, (2) provide the selective pressures to change the distribution of antibodies in the bloodstream, that is, cause the *differential* reproduction of B-cells, leading to a first selection process. At this point,<sup>8</sup> the pathogen population has caused a new distribution of prior B-cells in the bloodstream, call it the "infection phase one population." The structure of this population results from the increased reproduction of B-cells, the action of the mechanism that selects against self-binding (or anti-self) antibodies and the mechanism that selects for the antigen-binding antibodies. The rise in B-cell reproduction caused by the pathogens present in the current pathogen population also (3) causes an increase in mutation rate of the 200 or so gene segments that code for parts of the antibody molecule, which in turn (4) causes an increase in antibodies whose structure closely matches that of the pathogens' antigen, which (5) makes the immune system able to (a) home in on the specific variety of pathogen it currently faces and (b) eventually track its evolution (that is, the pathogen's adaptation to the immune systems current distribution of antibodies). The distribution of prior B-cells in the bloodstream is under selective pressure and the immune system's complex response coupled with the system's interaction with the pathogen population currently in the bloodstream results in the post B-cells, which, to make things simple, can be viewed as the cell population with which the organism will face its next infection, that is, as the next

---

8. Of course, this is not a real point in time, since any division between this process and the next is highly artificial and only useful for exposition purposes.

prior B-cell population (although nothing prevents selection process to act in between infections).

The processes described above are a good implementation of the exploration/exploitation trade-off in optimization algorithms. To successfully explore a space (searching for the set of parameters that would optimize its objective function), good optimization algorithms combine an exploration and an exploitation strategy. An exploitation strategy is a strategy that can maximize the algorithm's coverage of the search space. There are various ways to do this. One is to have an algorithm that can change its parameters quickly and substantially. This can be done in a genetic algorithm by having a very high mutation rate or by having high temperature in a simulated annealing algorithm (such as a Hopfield net). Another is to multiply the number of points in the environment that are simultaneously explored. This is done in ACO (ant colony optimization) algorithms by multiplying the number of "ants" that simultaneously explore the environment and in the immune system by the initial burst of 10,000 different specificities. However, an algorithm or set of algorithms that keeps exploration at such a high level won't be able to settle on the solution as it quickly loses (through variation) any solution it finds. At some point, variation needs to be toned down: in a simulated annealing algorithm temperature is lowered; in ant colony optimization, pheromones start driving the algorithm, thereby reducing the "freedom" of individual ants. In the immune system, it is the negative feedback loop between the number of pathogens in the bloodstream and the activity of B-cells that does the trick.

Enough has been said, we believe, to show the value of Darwin's way. Criticisms would abound if one were to try to greedily reduce the complex interplay of selection processes at work to set up and maintain a healthy immune system to one single algorithm forced on the domain because of its value in another. The lesson drawn from this body of work is that one should avoid dogmatism concerning the EAs and be ready to study them carefully (with an open mind) in any particular domain. This is a cautionary lesson, and it does not directly affect the benefits expected from the adoption of Memetics-D. Indeed, the revolution in our self-conception will be that much greater if we discover that culture is the result of many selection algorithms, operating at different time-scales on different units of selection and with different logical structures (see Dennett, 1995, for the logical structure of an algorithm). Any increased consilience with the rest of science accrued by Memetics-D will be surpassed by models that better reflect the complex nature of and possible interplay between selection algorithms. And finally, the more intricate the selection processes (and their interplay) at work in culture, the more space Memetics will have to capture the actual process that drives our acceptance of beliefs as knowledge.

### 3. Memes and Their Environment

The immune system is not the only place to look for inspiration for a reinvigorated Memetics. Genomics is not a bad place either. Indeed, the sequencing of the genome (the complete set of DNA within a single cell of an organism) of many organisms (including bacteria like *E. coli*, insects like fruit flies or *C. elegans*, plants like rice or *A. thaliana*, and animals like chickens or humans) and the analysis of its structure as well as the function(s) of its elements have provided the seeds for a revolution in the way we conceptualize the gene. The discovery of phenomena like alternative splicing, transplicing, frameshift, RNA editing, and others have made apparent the “limitations of the purely structural, sequence-based definition of the gene” at the center of the molecular conception of the gene (Griffiths & Stotz, 2006). Indeed, some argue that in these “postgenomics” days, DNA has lost its status as the “master molecule” that orchestrates the dance of life: it is now *one* participant of that dance, on the same level as strings of RNA, proteins, and factors that have traditionally been regarded as part of the environment. This has led some to the rejection of (molecular) gene-centrism, -realism, or -monism, which, in return, has led some philosophers of biology (and some biologists as well) to defend various forms of (molecular) gene skepticism, that is, deflationism (Falk, 2010; Griffiths & Stotz, 2006); pluralism (Moss, 2002; Griffiths & Neumann-Held, 1999); or outright eliminativism (Fox-Keller & Harel, 2007). Dennett is well aware of the reigning atmosphere of skepticism concerning the concept of gene. He writes, for instance, “Yes, of course, there are sequences of nucleotides on DNA molecules, but does the concept of gene actually succeed (in any of its rival formulations) in finding a perspicuous rendering of the important patterns amidst all the molecular complexity? If so, there are genes; if not, then genes will in due course get thrown on the trash heap of science along with phlogiston and the ether, no matter how robust and obviously existing they seem to us today” (Dennett, 2009, p. 2). But he does not seem worried too much by the change of paradigm that genomics could induce. This is because his concept of memes is like George Williams’s concept of genes: it is defined in terms of the information that a structure carries, not in terms of the material constituents of the structure.<sup>9</sup>

We do not contest this interpretation of what genes and memes are, but we think that Dennett may be overly optimistic if he thinks that the genomics

---

9. Dennett (2009) compares genes and memes to words, which “have physical ‘tokens’ (composed of uttered or heard phonemes, seen in trails of ink or glass tubes of excited neon or grooves carved in marble), . . . but these tokens are relatively superficial part or aspect of these remarkable information structures” (p. 4; our emphasis).



revolution will leave the concept of “gene as information” unscathed. This is what we want to show in the next subsection 3.1. This should also have an impact on the concept of “meme as information.” But instead of signifying that the meme concept should be placed on the trash heap of science, we think it means that the concept should be deflated, and we see in this deflated concept a better tool for addressing some long-standing objections to Memetics-D.

### 3.1 *The New Old Replicators: The Trendy New Genes*

To understand the revolution provoked by genomics, one should first try to spell out what the traditional conception of gene is. The concept of gene that is threatened by genomics is the traditional molecular concept of gene inherited from Crick and Watson. This concept is captured by the idea that genes are “blueprints” or “programs” for building proteins. Griffiths and Stotz (2013, p. 40) summarize this concept thusly:

[T]he linear sequence of nucleotides in a segment of a DNA molecule specifies the linear sequence of nucleotides in an RNA molecule, and that molecule in turn determines the linear sequence of amino acids in a protein through ‘information specificity’—that is, via the genetic code

The idea is thus that there is some sort of “colinearity” between the DNA sequence and its products (the proteins), such that one could say that the DNA sequences contains the “image” of its products (an idea which could be dubbed “molecular preformationism”; Griffiths & Stotz, 2013, p. 101; cf., Shapiro, 2009). According to Griffiths and Stotz (2006, p. 512), this gene concept “has been derived from work on a limited range of organisms: prokaryotes and bacteriophages” (cf., Gerstein, Bruce, & Rozowsky, 2007, p. 671). Work on more complex organisms has favored a different picture (as with the case of the immune system, the moral might be that one should avoid settling too quickly in favor of one picture of EA or of the gene, as Mother Nature may have used various ways to reach her goals; cf., Godfrey-Smith this volume). This is the picture we want to present. But first, a bit of background:

- (1) Contrary to what was expected, the human genome turned out to contain a rather small number of protein-coding genes compared to that of other, much simpler, organisms. According to one estimate, the human genome contains 21,000 protein-coding genes (Gerstein et al., 2007, p. 669) while rice plants have around 38,000 of such genes (International Rice Genome Sequencing Project, 2005).



- (2) The percentage of protein-coding genes in the human genome is around 1.5%. Of the rest of the genome, 25% consists of introns and 50% of transposable elements and *pseudogenes*. It appears that there is no correlation between the percentage of protein-coding genes and an organism's complexity, though there is one between the ratio of transcribed, but non-coding DNA, to coding DNA (Griffiths & Stotz, 2013, p. 69).
- (3) In spite of its relatively small number of protein-coding genes (by comparison to more simple organisms), the human proteome (the sum total of the proteins produced by an organism) is estimated to contain one million proteins (Griffiths & Stotz, 2013, p. 70).

These facts have led researchers to conclude that posttranscriptional mechanisms must account for the number of proteins in human (this is what projects like ENCODE are trying to do, see ENCODE Project Consortium, 2012; Morange, 2014). The list of such mechanisms would be long (see Shapiro, 2009), but for the purpose of this paper, a sample of those should do.

- 1) *Alternative splicing*: it has been known since the 1970s that pre-mRNA is processed by cutting large non-coding sequences, call introns, while the remaining parts, call exons, are spliced together to form the final mRNA transcript (a process called "splicing"). Later, it was discovered that alternative versions of mature mRNA transcripts can result from the cutting and joining of different combinations of exons. In other words, one genetic locus can code for multiple different mRNA transcripts that code for different, but structurally related proteins (isoforms). This process is called "alternative splicing." The number of annotated alternative isoforms by locus is about 5.4 (Gerstein et al., 2007, p. 673), though Fox Keller and Harel (2007, p. 1) suggest that thousands of such transcripts could be formed from the same sequence of DNA. Alternative splicing requires a mechanism to specify which splicing should take place. Griffiths and Stotz (2013, p. 89) mention three mechanisms that can play such a role: "the synthesis of new splicing proteins by special regulator genes, the activation of splicing proteins through phosphorylation, and the movement of splicing regulatory proteins into the nucleus." According to Bradbury (2005), "the combinatorial mechanism for the control of alternative splicing ... *could allow cells to adjust splicing outcome* (and consequently which proteins they express) rapidly in response *to intra-cellular or extra-cellular cues* [our emphasis], as well to contribute to the generation of protein diversity" (quoted by Griffiths & Stotz, 2013, p. 90). For instance, Zang et al. (2015) describe how alternative splicing (controlled by

the exosome in response to changes in ambient temperature) plays a role in the regulation of circadian clocks in various species.

- 2) *Transplicing and RNA editing*: “Transplicing” is similar to splicing, but it joins pieces of two different transcripts, that can come from the same gene (allowing the inclusion of multiple copies of the same exons), or from opposite DNA strands or even from different chromosomes. Gerstein and his colleagues (2007, p. 672) describe a form of transplicing, called “tandem chimerism,” “where two consecutive genes are transcribed into a single RNA. The translation (after splicing) of such RNAs can lead to a new, fused protein, having parts from both original proteins.” “RNA editing” is a process through which some changes are made to specific nucleotide sequences. Such changes (the number of which can be controlled by metabolic or hormonal variations) can include insertion, deletion and base substitution (Sowden & Smith, 2005). As Griffiths and Stotz (2013, p. 94) remark, while transplicing “*scrambles* the order of the primary DNA sequence, RNA editing *changes* the primary sequence of mRNA during or after its transcription via the site-specific insertion, deletion or substitution of one of the four nucleotides.” It creates what have been dubbed “cryptogenes,” because this process leads to gene products whose image cannot be found in the DNA.
- 3) *Frameshift*: there are cases when pre-mRNA can be alternatively spliced to generate an mRNA with a frameshift in the protein sequence (that is, transcription will start from a different nucleotide, giving rise to two mRNA transcripts that have coding sequences in common, but different protein products). Programmed frameshifts (as opposed to accidental ones) are caused by *cis*-regulatory elements (i.e., elements in the mRNA sequence, like slippery sequences or pseudo-knots and stem-loops) or *trans*-acting factors ones (like proteins, micro-RNA or antibiotics). As Caliskan et al. (2015) put it: “Programmed frameshifting increases the coding potential of the genome and is often used to expand the variability of cellular proteome, *adapt to changing environments* [our emphasis], or ensure defined stoichiometry of protein products” (p. 265).
- 4) *Micro-RNA*: some stretches of DNA (continuous or not) are transcribed into pre-mRNA molecules that will not be translated into mature m-RNA, but that are involved in regulation of translation or post-translational events. Such RNA molecules, formed from stretches comprising about 21 nucleotides, are thought to be involved in target gene regulation and epigenetic silencing (Fox-Keller & Harel, 2007, p. 6), their dysregulation is known to play an important part in disease like cancer.
- 5) *DNA methylation and histones modification*: two processes are thought to contribute to the “epigenomic code” (Turner, 2007): DNA methylation

and histone modification. The first process refers to the addition of methyl groups to a segment of DNA (CpG dinucleotides) that acts to repress gene activity by blocking the binding of transcription factors, thereby silencing transcription. The second process refers to a combination of different modifications to the “tails” of histones (DNA is wrapped around a core of eight histone protein to form the nucleosome). These histones “protrude through the DNA and are exposed on the nucleosome surface, where they are subject to an enormous range of enzyme-catalyzed modification of specific amino-acid side chains the modifications decorate the nucleosome surface with an array of chemical information” [Turner, 2007, p. 2]) which alter the activity of genes via structural changes to the DNA molecule.<sup>10</sup> As Turner (2007; our emphasis) puts it, “patterns of histone modifications associated with ongoing transcription can change rapidly and cyclically *in response to external stimulation*. In this context, *histone modifications can be considered the endpoints, on chromatin, of cellular signaling pathways and a mechanism through which the genome can respond to the environmental stimuli* [our emphasis]” (pp. 2–3.) Both methylation and histone modification are thought to form an *epigenomic code*, a series of information that changes genomic functions (transcription, replication and repair, but also cellular memory) in response to internal and external environment.

We wish to draw two lessons from the recent genomic research. (1) In the process that leads from DNA to its products, the co-linearity of DNA can be profoundly disturbed by various mechanisms; “the linear sequence of a gene product is rarely specified or determined by its DNA sequence alone” (Griffiths & Stotz, 2006, p. 509). (2) Maybe more importantly, the informational flow is not unidirectional (going for the DNA to its products) but is the result of many different mechanisms that use DNA sequences in different ways in order to respond in a flexible fashion to the environment. To put it differently:

specific difference-making RNA and protein factors need to be recruited or activated by external inducers. These molecules undergo crucial changes in shape in response to these signals, which render them active and impose their causal specificity. *One can indeed say that environmental signals are the drivers of gene expression* [our emphasis]. The significance of this is that these molecules relay specific difference making (instructive) environmental information to the genome (Griffiths & Stotz, 2013, p. 101).

---

10. Some go as far as talking of a “histone code” (see Shapiro, 2013, p. 294; Griffiths & Stotz, 2013, p. 70).

The genome should not be thought as the unique source of information (as when it was thought to be a blueprint or a program), but as something utilized by the cell in response to other information coming from the environment.

As we will see in the rest of this section, what is true for genes (that they are what the cell or the organism is doing with its genome) might also be true for memes. Memes would be what an organism does with its “memome” (we propose this term to designate the complete set of informational structures that populate an organism’s brain as a result of memetic selection and which are used by it to create its products, that is, behaviors, ideas, artefacts, etc.). As we said earlier, thinking of memes along these lines might not only be more biologically valid, it could also help to counter some serious objections to Memetics. We will consider one in the next sub-section.

### 3.2 *Maurice Bloch’s Objection: Memes Are Not the Units of Culture*

Over the years, Dennett has answered many objections to Memetics-D, including objections to the Lamarckian character of cultural evolution, to the fact that cultural evolution is sometimes horizontal rather than vertical, and to the intentional character of cultural innovation (see for instance, his 1999). In this section, we will present an objection raised by the anthropologist Maurice Bloch that we find particularly interesting (and to which Dennett, to our knowledge, has yet to provide an answer). We think that adopting a conception of memes inspired by the new conception of genes we just presented might help proponents of Memetics-D to answer this objection (by adopting Memetics-P). Once again, it would be foolish of us (and not in the spirit of Darwin’s Darwinism) to propose a replacement structure (a type of meme) for the one greedily chosen by early adopters of Memetics-C. But we consider what happens to the gene’s concept as an eye-popper: as biological systems got more complicated, they evolved different ways to use the information contained in the genome. It is possible that the same happened with the brain structures that constitute the “meme’s ‘cryptic’ form” (Haig, 2006, p. 1). Thus, though we think that it is quite possible that some memes are more like the old-fashioned genes, others might behave more like the new ones we described in the previous sections. We will focus on the latter in the following.

In his “A Well-Disposed Social Anthropologist’s Problems with Memes” (2000; see also his 1998, 2005, and 2012), Bloch describes a debate that took place in anthropology at the turn of the century concerning the best way to characterize culture. He argues that the same criticisms that were leveled back then against diffusionist’s theories can be made against Memetics-D, and he suggests

that “memeticians” should pay more attention to anthropology to avoid repeating past mistakes.

Diffusionism was, like Memetics-D, a theory aimed at understanding the patterns of distribution of cultural items in human culture across the world. It emphasized the necessity to understand cultural items as largely independent, explicit, and atomistic elements that could spread from culture to culture. Diffusion, in this context, could be understood as the spread of a discrete cultural item from its place of origin to other places by means of contact (either communication or imitation).

Diffusionism was criticized for having a notion of culture (as a repository of discrete atomistic cultural items) that “was too decontextualized from the *practice* of ordinary life” (Bloch, 2000, p. 199). According to Bloch, the same criticism can be leveled against Memetics. Bloch rejects the intellectualist’s view of culture, which he sees at the core of Memetics-D, that reduces it to a set of *explicit* representations (words, definitions, rules, representations or classifications). To be clear, he is not saying there are no representations underlying these practices, nor that these representations are not transmitted between individuals. He is also not implying that the receivers of these representations do not store them in their minds. Rather he stresses “that this knowledge is often *implicit*; that it does not exist in a vacuum. As a result, it is so intimately implicated in action and interaction that it only exists as part of a whole, only one aspect of which is purely intellectual in character. *To represent culture as a collection of bits of information is thus to forget that most of the time it cannot be separated from practices*, to which it relates in a number of fundamentally different ways” (2000, 200; our emphasis). In a nutshell, culture is formed of explicit and implicit representations. The latter (which form most of culture; 1998, p. 14) are not independent of practices (there are no pure intellectual things); they have an immediate (or at least, more immediate than explicit representations) link with action.

Throughout his work, Bloch denounces the dominant school in anthropology (and by the same means, Memetics-D) that rests on a language-like model of culture, with discrete explicit and disincarnate representations, accessible directly through introspection to members of a culture. Bloch rather sees culture as formed by “lived-in models, that is, models based as much in experience, practice, sight and sensation as in language” (Bloch, 1998, 25). More precisely, he conceives of it as consisting of a number of mental (or cultural) models or scripts or schemata partly accessible to language, but also partly visual, sensual and linked to action (Bloch, 1998, 26). As he says, “such a schema permits us not only to recognise an office—the various elements which make up its furniture—but *also to know how to react toward it in an appropriate manner* holding such a schema enables the individual to recognise not just a particular office but

all the occurrences of what could be an office and *to act according to all the possible requirements of this category in a quasi-automatic fashion, without paying much conscious attention to the actions which an office is likely to entail for them*" (Bloch, 1998, 45; our emphasis).

Bloch also observes that this implicit knowledge, on which ordinary practices are based, "is often quite at odds with explicit beliefs declared by the people studied or by those who study them (anthropologists for instance), especially when these base themselves principally on the declarations and symbolic aspects of the behavior of those they observe. With such an attitude, British anthropologists [including himself] see culture as existing on many levels, learnt implicitly or explicitly in a great variety of ways" (Bloch 2000, 200). If this is the case, culture is complex, in the sense that it groups different kinds or types of knowledge (some propositional, some not), stored and learned in different ways and variably accessible to consciousness.

As we will argue in section 3.3, Bloch's objection does not constitute a strike at the heart of Memetics-C. In shedding light on an aspect of culture that Memetics-D has neglected, we think that it brings about a refinement that should be welcomed. But, accepting this refinement might require the acceptance of a new framework to understand the meme, one along the lines of the new view of genes we addressed in section 2.1. To be entirely clear about what the refinement is and why it requires a change in the conception of the memes, we will present an example, taken from the literature in social cognitive psychology.

### 3.3 *The Multiple Memory Systems Model of Stereotype Representations*

Discrimination, conflicts, and disparities based on race, gender, ability, identity or body styles, etc. are not on the decline. Open your daily paper and you will read about multiple instances of discrimination and prejudice everywhere in your society as well as in others more distant. But, at least in some countries, norms now favor the open expression of more egalitarian ideas and chastise the open expression of prejudice. Old-fashioned, overt racism is now frowned upon in many parts of society, enough so that most people would not dare to express such attitudes. As Amodio (2014a) puts it, this situation has led prejudices to "go underground," that is, it has led them to operate covertly making them more difficult to detect. People studying prejudice have had to adjust: they have gone, so to speak, "undercover." For instance, they have modified their self-report measuring scale so it would not be obvious to subjects that the test is tapping into their prejudices anymore (i.e., the modern racism scale). Other measures tap subtler expressions of bias that are thought not to be controllable and/or accessible

to consciousness: they measure what have been called “Implicit Attitudes” or “Implicit Biases.” In the following, we will focus our attention on what is measured by the latter.

To tap into implicit prejudicial attitudes (not directly accessible to the experimenter either due to the subject’s social desirability concerns, or because these attitudes are unconscious) social psychologists have used “indirect” methods such as semantic priming, implicit association tests (IAT), affect misattribution procedure, etc. (for a review of these methods, see Nosek et al. 2011).<sup>11</sup> It has been shown that these implicit methods outperform explicit ones in predicting behaviors, choices, or judgments in socially sensitive domains (Greenwald et al. 2009; Pearson et al. 2009; Rudman & Ashmore, 2007), though there are now debates about the size of the effects discovered and their importance (Greenwald et al., 2015; Oswald et al. 2013)<sup>12</sup>. Recent research on the developmental course of implicit social cognition (Olson & Dunham, 2010) has shown that the content of (at least some of) the attitudes is learned early and remains relatively stable through development (contrary to explicit social cognition that changes quite radically from childhood to adulthood). This is taken to be proof that these attitudes are not produced by a slow-learning system culling the environment for regularities, but by early dispositions to learn about social group (some sorts of prepared learning; see Machery & Faucher, 2005; Moya and Boyd, ms).

There are many distinct ways in psychology to models implicit bias (see Payne & Gawronski, 2010). We will focus on one of these models here: the “memory systems model” developed by Amodio and his colleagues (Amodio & Ratner, 2011; Amodio, 2014a, b; for similar models, though based on different principles, see Huebner 2016 and Van Bavel et al. 2012). We will present their most recent proposal and we will complement it with a tentative framework to understand how it could be implemented in “the wild.”

It is customary, in the literature on implicit attitude, to speak as if the mind were dual in that it comprises explicit and implicit processes (Frankish, 2010).

---

11. The term “indirect” is used to refer to the features of measurement procedures that “provide indicators of psychological attributes (e.g., attitudes) without having to ask participants to verbally report the desired information” (Payne & Gawronski, 2010, p. 4).

12. There is also a debate concerning the nature of the associations detected by the IAT. Some (Nosek & Hansen, 2008) argue that these associations are all “personal” (that is, they measure attitudes that shape our behaviour, instead of just “inert knowledge” of associations prevalent in our culture); others (Olson et al., 2009) rather think that a larger subset of them are “extra-personal,” while a small subset is personal (for this reason, they devised a different version of the IAT to tap into them). For the sake of space, we’ll leave that matter aside, since it is enough for us that at least some of the associations detected by IAT are personal and are acquired through a cultural process of transmission, a thesis not contested by anyone in this debate.



For instance, Gaworonski and Bodenhausen (2006) suggest that the distinction between explicit and implicit rests on a distinction between propositional and associative processes (accordingly, they dubbed this model, the “associative-propositional model”). It is this image of the implicit mind being constituted by a unified domain of processes that has been challenged by Amodio and his colleagues “Memory Systems Model” (MSM). Taking their cue from the cognitive neuroscience of memory<sup>13</sup> that has distinguished several forms of non-declarative memory<sup>14</sup> and identified different locations in the brain as their substrates, Amodio and Ratner (2011) propose that implicit social cognition should be thought of along the same lines. They propose considering a subset of the systems that could underlie implicit social cognition. These systems are semantic associative memory (in charge of forming associations between cognitive concepts, for instance between a target group and a property—Asians and mathematics), fear memory (in charge of forming associations between a target group and an affective response—White and fear) and instrumental learning based memory (in charge of forming associations between a target group and an action tendency—Black and avoidance). It is postulated that each system has different properties (that is, racial biases are learned, stored and expressed differently by each system) and has different neural substrates (neocortex—semantic memory; amygdala—fear memory; striatum and basal ganglia—instrumental memory). Research has already begun to confirm the MSM (Olsson, Ebert, Banaji, & Phelps, 2005) and has led to important insights concerning the acquisition along with potential ways of getting rid of some of types of associations (Shurick et al., 2012).

While Amodio and his colleagues have emphasized distinctions between different forms of implicit memory systems, it is believed that these systems typically work in concert to produce social behavior.<sup>15</sup> Though they are not keen on furnishing a model of the interaction between these systems, such a model should be provided to understand social cognition “in the wild.” We introduce such a model of implicit racial prejudice here: the “socially situated and embedded theory of concepts.” In introducing this model, we will be following Niedenthal and

---

13. For an overview of this type of work see Squire and Wixted (2011).

14. Declarative memory refers to consciously accessible memory, while non-declarative memory refers to memory that is not consciously accessible, but that could be expressed in behavioral tasks.

15. As they themselves admit: “. . .[i]mplicit attitudes accessed by sequential priming tasks may reflect a combination of semantic associations (e.g., with Good vs. Bad concepts), threat- or reward-related affective associations, and instrumental associations (e.g., reinforced and habitual actions)” (Amodio & Ratner, 2011, p. 145).



his colleagues' (Niedenthal et al. 2005; but also Barsalou, 2008) description of the Perceptual Symbol System (PSS) theory.

A situated and embodied theory of concepts asserts that concepts use "partial reactivations of states in sensory, motor, and affective systems to do their jobs ... The brain captures modality-specific states during perception, action and interoception and then reinstates parts of the same states to represent knowledge when needed" (Niedenthal, 2007, p. 1003). Representations are then modal (they are realized by modality-related re-activations), not modular (in that the conceptual system is not autonomous from other systems), contextualized (in that conceptualizations are linked to situations or contexts) and fluid (in that concepts deliver different packages of information in different contexts or situations).

Among the central constructs of PSS are the concepts of "simulator," "simulations" and "situated concepts." We want to introduce these concepts through two examples. Let's take first the concept of "rage." According to PSS (see Niedenthal, 2007 for instance), it is represented in a multimodal fashion, including sensory, motivational, motor and somato-sensory features which are typical of episodes of rage. It also contains or it is linked to information about situations in which rage is typically produced.

All these associated representations form a "simulator" which can create "on the fly" representations adapted to particular instances of a category ("simulations").<sup>16</sup> For example, in some contexts, our mental representation of "rage" might include a strong visceral reaction, but this might not be included in others. Thus, "situated conceptualizations" of rage will include only properties that are contextually relevant. For instance, when thinking about "rage" in the context of "road rage," one will represent a loss of control, swearing, violent actions, etc.; while thinking about the "rage" of an athlete who wants win an event, one's depiction will include looks of determination on their face, concentration, stress, etc.<sup>17</sup> Conceptualizations are thus *situated* in that a simulator will deliver a different package of information (a simulation) in different situations. As Barsalou (2008, p. 144) puts it: "From this perspective [the perspective of PSS], a concept is neither a static database nor a single abstraction. Instead, it is an ability or competence to produce specialized category representations that support goal pursuit

---

16. As Niedenthal and colleagues put it: "According to PSS, the simulation process is highly dynamic and context dependent. ... Depending on the current state of the simulator, the current state of associated simulators, the current state of broader cognitive processing, and so forth, a unique simulation results" (2005, p. 196).

17. Wilson-Mendenhall et al. (2013) suggest that emotional experience might be situated as well. They compared cerebral imagery of subjects imagining fear of speaking in public and fear for one's physical safety and found shared patterns of activity in multi-modal sensory areas across both situations, but also unique activation patterns for each situation.

in the current setting, where each specialized representation is akin to an instruction manual for interacting with a particular category member.” Over time, the situated conceptualization comes to mind automatically when a particular situation is detected. One might think that these automatic situated conceptualizations are exactly what social psychologists are referring to when they talk about “implicit cognition.”

Let’s return now to implicit prejudice. As was seen earlier, different memory systems are involved in stocking information about a category: semantic, affective, instrumental. PSS invites us to think that in real or “wild cognition” these different systems are integrated (and dynamical; see Van Bavel et al., 2012). Thus for PSS, our concept (simulator) of a social group (a race or a gender) will contain information not only about characteristic traits of the group (“Xs are good dancers but bad curlers”), but also auditory and visual representations as well as characteristic affective reactions and motor reactions toward members of it. This way of thinking about concepts has the advantage of explaining the link between the concept and action in a way that is not otherwise explained, or not explained at all, by an abstract, amodal view of stereotypical representations. For instance, studies (e.g., Chen & Bargh, 1999) have documented that if one is required to pull a lever toward or away from oneself to report the valence of a word, the movement used will affect the speed of response: the response is faster for a positive word if the lever motion is toward oneself, and slower if the lever is pushed away (and vice versa for negative words; see also Paladino & Castelli [2008] and Slepian et al. [2012]). This is supposed to show that there are associations between “self” and “positive” and “other” and “negative.” But it also shows that this association is not (only) abstract; it also comprises a motoric element (and is probably linked with instrumental memory). It also explains why damage to the amygdala does not affect IAT, but only physiological responses. In this case, the affective part of the concept is missing, but other aspects are present—and it would be interesting to know which ones. The reason this question is important is because PSS distinguishes between different ways of representing concepts.<sup>18</sup> This is made clear by a distinction introduced by Niedenthal and colleagues (2005, p. 199) between *shallow* and *deep* processing. According to them, a task needs deep processing when it requires a simulation for its completion. When it does not, that is, when one can perform a task by only using word-level representation, it is said to

---

18. Huebner (2016) argues that the properties attributed to implicit biases result from the operation of different fast and automatic learning systems. As he puts it, these different learning systems cast their votes concerning the course of action to adopt. Some unexplained effects observed in the implicit attitude literature (for instance, the increase in response latency in certain cases) could be explained by the fact that these systems are not always synchronized (i.e., they cast contradictory votes).

be shallow. So one important question when one performs an IAT for instance is: what am I tapping into? Is this association I am detecting the reflection of a “superficial,” purely linguistic, representation or is it the reflection of something deeper—a partial simulation—that could have direct impact on future action? For instance, it has been shown in consumer attitudes and behavior research that “implicit emotions evoked by consumer products predicted participants’ buying intentions significantly better than their general implicit attitudes toward those products. Interestingly, this finding nicely parallels intergroup emotions literature which also makes the case that the specific emotions people feel toward various outgroups predict their action tendencies better than global evaluations, probably because emotions are more tightly linked to specific goal-directed action tendencies compared to global evaluations” (Dasgupta, 2010, 55).

Finally, the PSS conceptualization fits well with a theme of social cognition that is gaining momentum nowadays: the malleability of implicit attitudes. This malleability could be explained by the fact that “[h]uman cognitive systems produce situated versions of concepts that have context-specific functions rather than activating the same, context-independent configuration in every situation” (Smith & Semin 2007, p. 134). As was seen earlier, PSS predicts that simulators will deliver different packages of information depending on context. For instance, if an individual is angry or afraid for irrelevant reasons, negative stereotypes will be more readily applied to out-group members than when that individual is in a neutral or happy state (Smith & Semin 2007, p. 133; cf., Feldman and Simmons 2015 and Feldman and Wormwood 2015). Similarly, in a dark alley or when cues of vulnerability to disease are made salient, specific prejudices concerning some group are activated, while others, applicable to the same group, are not. For instance, Schaller and Conway (2004) report that “[d]arkness amplified prejudicial beliefs about danger-relevant traits (trustworthiness and hostility), but did not much affect beliefs about equally-derogatory traits less relevant to danger” (155). Dasgupta and colleagues (2009, 589) report similar findings (this time, specifically on implicit attitudes), showing that “[r]ather than serving as a general warning that orient perceivers to generic dangers thereby increasing bias against any outgroup, emotions exacerbate bias only when the feeling warns of a specific threat that is directly applicable to the outgroup being appraised.”

Results such as those presented in the previous paragraph have led many to conclude that implicit attitudes are not static representation, but dynamical ones. As Payne and Gawronski put it: “In the eyes of construction theorists, the high malleability of indirect measurement scores confirmed their assumption that contexts influence what information is used to construct an attitude from one moment to the next, and that these principles apply equally to direct and indirect measurement procedures. In fact, the very idea that indirect measurement

procedures would assess rigid ‘things’ in memory independent of the context was seen as ill founded.” (2010, 6).

### 3.4 *What Should Memetics-D Do?*

Bloch objects to Memetics-D that it wrongly presupposes that cultural items have to be (1) explicit and amodal, (2) propositional, (3) discrete and atomistic. As he argues, the anthropologist investigating another culture faces items that are (1) implicit and modal, (2) associative and (3) rather fuzzy and holistic. As we understand him, Bloch is claiming that it would be a mistake, in the process of trying to understand culture and cultural changes, to focus exclusively or primarily on explicit representations. But it would be a mistake to focus exclusively on implicit representations as well. It seems that both types of representations, with their respective properties, are parts of culture. It is quite possible that Memetics-D has given the impression of paying more attention to explicit representations, but it is in no way committed to doing so. For instance, Dennett has mentioned “fashion” amongst the things that Memetics-D might explain (2006b). A particular fashion trend might not be something that can completely be represented explicitly; it can include some sort of general way of thinking about how to dress, what to drink, what kind of activities or music are better than others. Think of being a hipster or a punk, for instance. It is something that is more than the words used to describe the style (“hipster,” “punk”), even more than some explicit non-verbal representations of the style in magazines (because, you can dress like one and not really be one). It looks more like the mental structures of the type Bloch alludes to when talking about culture. Memetics-D, which often talks about larger structures in terms of “memplexes,” might not put enough emphasis on the distinction between explicit and implicit representations. Bloch’s remarks might then be seen as a corrective to Memetics-D, not a strike at the heart of the project.

The work on implicit attitudes we reviewed in the previous sub-section illustrates Bloch’s position. It shows that part of what we call culture is implicit, not necessarily or easily accessible to subjects (so they might have problems articulating it). In the case of attitudes (but this is not necessarily true in every domain of cognition), the content of explicit and implicit attitudes might be quite different, so much so, that knowing what people explicitly believed might not be a very good predictor of what they will think, how they will judge a situation, how they will feel or act.

We think that adopting this model of culture will further Dennett’s goals in adopting Memetics-D. For instance, it definitively modifies our self-image. For instance, Bargh and Chartrand conclude that “most of a person’s everyday life is determined not by conscious intentions and deliberate choices but by

mental processes that are put into motion by features of the environment and that operate outside of conscious awareness and guidance” (Bargh & Chartrand, 1999, p. 462). As Bargh (1999) puts it, “contrary to what we think of the reasons that motivate our actions, a large part of our behavior is in the hands of a cognitive monster inside us that cannot be controlled.” If this is the case, “we” (that is, our conscious self) are not the ultimate spring of some of our actions. This is a conclusion that Dennett would be more than happy to subscribe to. As he wrote recently (but the theme goes at least back to his 1991): “All the work done by the imagined homunculus in the Cartesian Theater must be distributed around to various lesser agencies in the brain ... A lot of what *you* are, a lot of what you are doing and know about, springs from structures down there in the engine room, causing the action to happen” (2003, pp. 39, 49). But the work we reviewed is adding a twist to what Dennett says: the lesser agencies that determine our actions might be created on the fly, in response to features of a particular context. The content of the engine room would thus be more evanescent or unstable than we might have thought. As for consilience with other scientific disciplines, we think that adopting this model would be doubly beneficial for Memetics. First, it would increase the consilience of Memetics with cognitive psychology and neuroscience, which is something that has been called for by many (Plotkin, 2000; Sperber, 2000; Atran, 2001). Second, it would increase consilience with biological disciplines. The fact that memes are conceptualized as information which can be therefore multiply realized by different structures should not be understood as preventing Memetics from taking cues from recent research. More specifically, the lessons we draw about the concept of gene from the work in genomics can be imported to Memetics: like genetic structures, some memetic structures can be differentially activated depending of the context and deliver different information packages in different situations. Finally, adopting this model would also further Dennett’s last motivation for adopting Memetics-D. Indeed, the presence of contradictory attitudes helps to explain irrational behavior (for instance, the fact that some people discriminate despite their commitment to treat everybody equally). It also shows that some representations can be deeper than others, by which we mean that they can involve more than semantic knowledge, but also include an affective and behavioral aspect.

Finally, though space prevents us to develop this further, an interesting idea would be to study the processes of transmission and selection that operate on explicit and implicit attitudes (something similar to what we did in section 2 for antibodies).<sup>19</sup> As we said earlier, no one doubts that some of these attitudes

---

19. We thank Tad Zawidzki for suggesting us this idea.

are “caught” from one’s cultural environment. There are some indications that what we might find out by studying the transmission and selection of attitudes is that many different types of processes operating underlie them, with different logics (see, for instance, Olsson & Phelps, 2007; Huebner, 2016). Some of these processes will include loops into the environment (see for instance, Sundstrom, 2003, and Huebner [ms]). Others will include loops in the organism: for example transmission of some implicit attitudes might depend on the internal states of the individuals (for instance, their beliefs in essentialism [see Haslam et al., 2006] or their awareness of their social location. See Mahalingam, 2007). If this proves to be true, it would militate even more strongly for the adoption of a Memetics-P.

#### 4. Conclusion

We owe a great debt to Dennett’s efforts to have revived the Memetics-C project (in his particular version, Memetics-D). Only the future will tell if this project turns out to be a fruitful scientific endeavor (Dennett himself seems to be skeptical about that and we are as well). One thing is certain though: there is absolutely no chance for Memetics-C to reach a scientific status if it takes dogmatic positions concerning its main concepts (like the evolutionary algorithms or the structure of memes). As we have argued in this paper, Memetics-C should be attentive to what is going on in its disciplinary neighborhood if it wants to eventually qualify as a science. For this reason, we have argued that Memetics-D should give way to Memetics-P.

#### Acknowledgments

We would like to thank Bryce Huebner and Tad Zawizki for their numerous and generous comments on previous versions of this paper. L.F. would also like to thank André Rato for his precious assistance.

#### Works Cited

- Amodio, D. M. (2014a). Dual experiences, multiple processes: Looking beyond dualities for mechanisms of the mind. In J. S. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual process theories of the social mind* (pp. 560–576). New York, NY: Guilford Press.
- Amodio, D. M. (2014b). The neuroscience of prejudice and stereotyping. *Nature*, 15, 670–682.
- Amodio, D. M., & Ratner, K. G. (2011). A memory systems model of implicit social cognition. *Current Directions in Psychological Science*, 20, 143–148.

- Atran, S. (2001). The trouble with memes: Inference versus imitation in cultural creation. *Human Nature*, 12, 351–381.
- Bargh, J. A. (1999). The cognitive monster: The case against controllability of automatic stereotype effects. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 361–382). New York, NY: Guilford Press.
- Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, 54, 462–479.
- Barsalou, L. (2008). Situating concepts. In P. Robbins & M. Ayede (Eds.), *Cambridge handbook of situated cognition* (pp. 236–263). New York, NY: Cambridge University Press.
- Bloch, M. (1998). *How we think they think: Anthropological approaches to cognition, memory and literacy*. Boulder, CO: Westview Press.
- Bloch, M. (2000). A well-disposed social anthropologist's problems with memes. In R. Aunger (Ed.), *Darwinizing culture: The status of memetics as a science* (pp. 189–203). New York, NY: Oxford University Press.
- Bloch, M. (2005). *Essays in cultural transmission*. [Series]. London School of Economics: Monographs in Social Anthropology 75. Oxford, England: Berg.
- Bloch, M. (2012). *Anthropology and the cognitive challenge*. Cambridge, UK: Cambridge University Press.
- Bradbury, J. (2005). Alternative mRNA splicing: Control by combination. *PLoS Biology*, 3(11): e369. doi:10.1371/journal.pbio.0030406
- Breakey, A., Hinde, K., Vallengia, C., Sinofsky, A., & Ellison, P.T. (2015, January). Illness in breastfeeding infants relates to concentration of lactoferrin and secretory immunoglobulin A in mother's milk. Online publication. *Evolution, Medicine, and Public Health*, 2015(1), 21–31. doi:10.1093/emph/eov002
- Caliskan, N., Peske, F., & Rodina, M. V. (2015). Changed in translation: mRNA recoding by -1 programmed ribosomal frameshifting. *Trends in Biochemical Sciences*, 40, 265–274.
- Campbell, D. T. (1974). Unjustified variation and selective retention in scientific discovery. In F. J. Ayala & T. Dobzhansky, *Studies in the philosophy of biology* (pp. 139–161). London: MacMillan.
- Chen, M., & Bargh, J. (1999). Nonconscious approach and avoidance behavioral consequences of the automatic evaluation effect. *Personality and Social Psychology Bulletin*, 25, 215–224.
- Craver, C. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. New York, NY: Oxford University Press.
- Dasgupta, N. (2010). Implicit measures of social cognition: Common themes and unresolved questions. *Journal of Psychology*, 218, 54–57.
- Dasgupta, N., DeSteno, D., Williams, L. A., & Hunsinger, M. (2009). Fanning the flames of prejudice: The influence of specific incidental emotions on implicit prejudice. *Emotion*, 9, 585–591.
- Dawkins, R. (1976). *The selfish gene*. New York, NY: Oxford University Press.



- Dennett, D. C. (1990). Memes and the exploitation of imagination. *Journal of Aesthetics and Art Criticism*, 48(2), 127–135.
- Dennett, D. C. (1991). *Consciousness explained*. Boston, MA: Little, Brown.
- Dennett, D. C. (1995). *Darwin's dangerous idea: Evolution and the meanings of life*. New York, NY: Simon and Schuster.
- Dennett, D. C. (2000). Foreword. In R. Aunger (Ed.), *Darwinizing culture: The status of memetics as a science* (pp. vii–ix). New York, NY: Oxford University Press.
- Dennett, D. C. (2001). The evolution of culture. *The Monist*, 84, 305–324.
- Dennett, D. C. (2003). The self as a responding and responsible-artifact. *Annals of New York Academy of Sciences*, 1001, 39–50.
- Dennett, D. C. (2006, June). Religion's just a survival meme. *Science and Theology News online*, available at <https://ase.tufts.edu/cogstud/dennett/papers/ScienceTheologyNews.pdf>
- Dennett, D. C. (2006b). There aren't enough minds to house the population explosion of memes. Online publication. Edge.org. <https://edge.org/response-detail/11320>
- Dennett, D. C. (2009, August). The cultural evolution of words and other thinking tools. *Cold Spring Harbor Symposia on Quantitative Biology*. Online publication. <http://dx.doi.org/10.1101/sqb.2009.74.008>
- Dennett, D. C. (2011). Homunculi rule: Reflections on *Darwinian Populations and Natural Selection* by Peter Godfrey-Smith. *Biology and Philosophy*, 26, 475–488.
- Dennett, D. C. *Memes: Myths, misunderstandings and misgiving*.
- Depew, D. J., & Weber, B. H. (1995). *Darwinism evolving: Systems dynamics and the genealogy of natural selection*. Cambridge, MA: MIT Press.
- Dreyfus, H., & Dreyfus, S. (1986). *Mind over machine: The power of intuition and expertise in the era of the computer*. Oxford, UK: Blackwell.
- Edelman, G. (1987). *Neuronal Darwinism: The theory of neuronal group selection*. New York, NY: Basic Books.
- Falk, R. (2010). What is a gene? Revisited. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 41, 396–406.
- Faucher, L. (2016). Revisionism and moral responsibility. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy* (Vol. 2, pp. 115–144). New York, NY: Oxford University Press.
- Feldman Barrett, L., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews: Neuroscience*, 16, 419–429.
- Feldman Barrett, L., & Wormwood, J. (2015, April). When a gun is not a gun. *New York Times, Sunday Review* (consulted online, April 22, 2017: <https://www.nytimes.com/2015/04/19/opinion/sunday/when-a-gun-is-not-a-gun.html?smprod=nytcore-ipad&smid=nytcore-ipad-share>).
- Fernando, C., Szathmari, E., & Husbands, P. (2012). Selectionist and evolutionary approaches to brain function: A critical appraisal. *Frontiers in Computational Neuroscience*, 6(24). <http://dx.doi.org/10.3389/fncom.2012.00024>
- Fox-Keller, E., & Harel, D. (2007). Beyond the gene. *PLoS One*: e1231. doi:10.1371/journal.pone.0001231



- Franfurt, H. (1971). Freedom of will and the concept of person. *Journal of Philosophy*, 68, 5–20.
- Frankish, K. (2010). Dual-process and dual-system theories of reasoning. *Philosophy Compass*, 5, 914–926.
- Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, 132, 692–731.
- Gerstein, M. B., Bruce, C., & Rozowsky, J. S. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Research*, 17, 669–681.
- Godfrey-Smith, P. (2011). *Darwinian populations and natural selections*. New York, NY: Oxford University Press.
- Gould, S. J. (1996). *Full house: The spread of excellence from plato to darwin*. New York, NY: Harmony.
- Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the implicit association test can have societally large effects. *Journal of Personality and Social Psychology*, 108, 553–561.
- Greenwald, A. G., Poehlman, T. A., Uhlmann, E., & Banaji, M. R. (2009). Understanding and using the implicit association test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, 97, 17–41.
- Griffiths, P. E., & Neumann-Held, E. (1999). The many faces of the gene. *Bioscience*, 49, 656–662.
- Griffiths, P. E., & Stotz, K. (2006). Genes in the postgenomic era. *Theoretical Medicine and Bioethics*, 27, 499–521.
- Griffiths, P. E., & Stotz, K. (2013). *Genetics and philosophy: An introduction*. New York, NY: Oxford University Press.
- Haig, D. (2006). The gene meme. In A. Grafen & M. Ridley (Eds.), *Richard Dawkins: How a scientist changed the way we think* (pp. 50–65). New York, NY: Oxford University Press.
- Haslam, N., Bastian, B., Bain P., & Kashima, Y. (2006). Psychological essentialism, implicit theories, and intergroup relations. *Group Processes and Intergroup Relations*, 9(1), 63–76.
- Haugeland, J. (1985). *Artificial intelligence: The very idea*. Cambridge, MA: MIT Press.
- Heinrich, J. (2015). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton, NJ: Princeton University Press.
- Huebner, B. (2016). Implicit bias, reinforcement learning, and scaffolded moral cognition. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy* (Vol. 1, pp. 47–79). New York, NY: Oxford University Press.
- Huebner, B. The dangers of white spaces. [Unpublished manuscript].
- Hull, D. (1988). *Science as a process: An evolutionary account of the social and conceptual development of science*. Chicago, IL: University of Chicago Press.

- Hull, D., Langman, R. E., & Glenn, S. S. (2001). A general account of selection: Biology, immunology, and behavior. *Behavioral and Brain Sciences*, 24, 511–528.
- Hutter, T., Gimbert, C., Bouchard, F., & Lapointe, F.-J. (2015). Being human is a gut feeling. *Microbiome*, 3(9). <http://dx.doi.org/10.1186/s40168-015-0076-7>
- International Rice Genome Sequencing Project. (2005). The map-based sequence of the rice genome. *Nature*, 436, 793–800.
- Laland, K., & Brown, G. (2002). *Sense and nonsense: Evolutionary perspectives on human behavior*. New York, NY: Oxford University Press.
- Machery, E., & Faucher, L. (2005). Social construction and the concept of race. *Philosophy of Science*, 72, 1208–1219.
- Malhingham, R. (2007). Essentialism, power, and the representation of social categories: A folk sociology perspective. *Human Development*, 50, 300–319.
- Mold, J. E., Venkatasubrahmanyam, S., Burt, T. D., Michaëlsson, J., Rivera, J. M., Galkin, S. A., . . . McCune, J. M. (2010). Fetal and adult hematopoietic stem cells give rise to distinct T cell lineages in humans. *Science*, 330, 1695. <http://dx.doi.org/10.1126/science.1196509>
- Morange, M. (2014). Genome as a multipurpose structure built by evolution. *Perspectives in Biology and Medicine*, 57, 162–171.
- Moss, L. (2002). *What genes can't do*. Cambridge, MA: MIT Press.
- Moya, C., & Boyd, R. *Whence ethnic psychology? An evolutionary functionalist reframing of the debate*. [Unpublished manuscript].
- Niedenthal, P. M. (2007). Embodying emotion. *Science*, 316, 1002–1005.
- Niedenthal, P., Barsalou, L., Winkielman, P., Krauth-Gruber, S., & Ric, F. (2005). Embodiment in attitudes, social perception, and emotion. *Personality and Social Psychology Review*, 9, 184–211.
- Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2011). Implicit social cognition: From measures to mechanisms. *Trends in Cognitive Sciences*, 15, 152–159.
- Nosek, B., & Hansen, J. J. (2008). The associations in our heads belong to us: Searching for attitudes and knowledge in implicit evaluation. *Cognition and Emotion*, 22, 553–594.
- Olson, K. R., & Dunham, Y. D. (2010). The development of implicit social cognition. In B. Gawronski & K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 241–254). New York, NY: Guilford Press.
- Olson, M. A., Fazio, R. H., & Han, H. A. (2009). Conceptualizing personal and extra-personal associations. *Social Psychology and Personality Compass*, 3, 152–170.
- Olsson, A., Ebert, J., Banaji, M., & Phelps, E. A. (2005). The role of social groups in the persistence of learned fear. *Science*, 308, 785–797.
- Olsson, A., & Phelps, E. A. (2007). Social learning of fear. *Nature Neuroscience*, 10, 1095–1102.
- Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, 105, 171–192.

- Paladino, M.-P., & Castelli, L. (2008). On the immediate consequences of intergroup categorization: Activation of approach and avoidance motor behavior toward ingroup and outgroup members. *Personality and Social Psychology Bulletin*, 34, 755–768.
- Payne, K., & Gawronski, B. (2010). A history of implicit social cognition: Where is it coming from? where is it now? where is it going? In B. Gawronski & K. Payne (Eds.), *Handbook of implicit social cognition* (pp. 1–15). New York, NY: Guilford Press.
- Pearson, A., Dovidio, J., & Gaertner, S. (2009). The nature of contemporary prejudice: Insights from aversive racism. *Social and Personality Psychology Compass*, 3, 1–25.
- Plotkin, H. (2000). Culture and psychological mechanisms. In R. Aunger (Ed.), *Darwinizing culture: The status of memetics as a science* (pp. 70–82). New York, NY: Oxford University Press.
- Richerson, P. J., & Boyd, R. (2005). *Not by genes alone: How culture transformed human evolution*. Chicago, IL: Chicago University Press.
- Rudman, L., & Ashmore, R. D. (2007). Discrimination and the implicit association test. *Group Processes Intergroup Relations*, 10, 359–372.
- Rutherford, E. (1911). The scattering of  $\alpha$  and  $\beta$  particles by matter and the structure of the atom. *Philosophical Magazine Series 6*, 21, 669–688.
- Schaller, M., & Conway, L. G. (2004). The substance of prejudice: Biological- and social-evolutionary perspectives on cognition, culture, and the contents of stereotypical beliefs. In C. S. Crandall & M. Schaller (Eds.), *The social psychology of prejudice: Historical and contemporary issues* (pp. 149–164). Lawrence, KS: Lewinian Press.
- Shapiro, J. A. (2009). Revisiting the central dogma in the 21st century. *Annals of the New York Academy of Sciences*, 1178, 6–28.
- Shapiro, J. A. (2013). How life changes itself: The read-write (RW) genome. *Physics of Life Reviews*, 10, 287–323.
- Shurick, A. A., Hamilton, J. R., Harris, L. T., Roy, A. K., Gross, J. J., & Phelps, E. A. (2012). Durable effects of cognitive restructuring on constructed fear. *Emotion*, 12, 1393–1397.
- Slepian, M., Young, S., Rule, N., Weisbuch, M., & Ambady, N. (2012). Embodied impression formation: Social judgments and motor cues to approach and avoidance. *Social Cognition*, 30, 232–240.
- Smith, E. R., & Semin, G. R. (2007). Situated social cognition. *Current Directions in Psychological Science*, 16, 132–135.
- Sober, E. (1980). Evolution, population thinking, and essentialism. *Philosophy of Science*, 47, 350–383.
- Sowden, M. P., & Smith, H. C. (2005). RNA edition. *Encyclopedia of Life Sciences*, Wiley, 1–7. <http://dx.doi.org/10.1038/npg.els.0003836>.

- Sperber, D. (2000). As objection to the memetics approach to culture. In R. Aunger (Ed.), *Darwinizing culture: The status of memetics as a science* (pp. 163–174). New York, NY: Oxford University Press.
- Squire, L. R., & Wixted, J. T. (2011). The cognitive neuroscience of human memory since H. M. *Annual Review of Neuroscience*, 34, 259–288.
- Sundstrom, R. (2003). Race and place: Social space in the production of human kinds. *Philosophy and Geography*, 6, 83–95.
- The ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489, 57–74.
- Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19–136). New York, NY: Oxford University Press.
- Turner, B. M. (2007). Defining an epigenetic code. *Nature Cell Biology*, 9, 2–6.
- Van Bavel, J. J., Xiao, J., & Cunningham, W. A. (2012). Evaluation is a dynamic process: Moving beyond dual system models. *Social and Personality Psychology Compass*, 6, 438–454.
- Watson, G. (1975). Free agency. *Journal of Philosophy*, 72, 205–220.
- Wilson-Mendenhall, C., Feldman Barrett, L., & Barsalou, L. (2013). Situating emotional experience. *Frontiers in Human Neuroscience*. <http://dx.doi.org/10.3389/fnhum.2013.00764>
- Zhang, L., Wan, Y., Huang, G., Wang, D., Yu, X., Huang, G., Guo, J. (2015). The exosome controls alternative splicing by mediating the gene expression and assembly of the spliceosome complex. *Scientific Reports*, 5, 13403. <http://dx.doi.org/10.1038/srep13403>

## 9.2 REFLECTIONS ON LUC FAUCHER AND PIERRE POIRIER

Daniel C. Dennett

The “fairy tale of sorts” with which Faucher and Poirer (F & P) introduce their topic is nicely aimed, drawing attention from the outset to the long, continuous lineage that starts with “a molecule” and arrives (today) at “a new type of replicator” that “eventually gave the survival machines the ability to understand how the world works.” It is that unbroken lineage, unaided by skyhooks, that makes Darwin’s idea so great.<sup>1</sup>

F & P give an excellent account of the “expected benefits” of the Memetics approach, which motivates their concern for the problem they see looming for my version: I have been lured by a *greedy algorithm* into a suboptimal solution, settling for what now turns out to be, on their reading, an obsolete version of evolutionary theory. The powerful alliance between gene and meme, introduced by Dawkins and adopted by me, turned heads when it was first proposed, but it also provoked a flood of resistance, not just from traditional social scientists and humanists, who didn’t want to see a Darwinian invasion of their territory, but also from evolutionary theorists who took themselves to be moving beyond the Establishment into new, maybe even revolutionary, territory. Using the Darwinesque operation of the immune system as an instructive example, F & P note that there are many different EAs (evolutionary algorithms) and say that I shouldn’t tie myself to the famous favorite, Darwinian natural selection. They suggest that I could do Memetics a lot of good by attaching it in one way or another to the new evo-devo, “postgenomics” bandwagon(s) and loosening my grip on the gene. “Any increased consilience with the rest of science accrued by Memetics-D will be surpassed by models

---

1. The Kandinsky painting on the dust jacket of *BBB* is entitled *Verso Continuo*, continuous line.

that better reflect the complex nature of and interplay between selection algorithms” (p. 200).

F & P give a clear and sympathetic description of some of these iconoclastic innovations, suggesting how they undermine, supplant, or at least extend neo-Darwinian orthodoxy, and summarize: “The lesson drawn from this body of work is that one should avoid dogmatism concerning the EAs and be ready to study them carefully (with an open mind) in any particular domain” (p. 267).

I agree entirely. There is no place for dogmatism here; Darwin’s idea generates defeasible scientific hypotheses, not theology, even if the reverence with which the Master is cited by some authors makes it seem as if Darwinism was a religion, like creationism. And F & P are right that, on the whole, I have sided with orthodoxy in my own expositions of Darwinian themes, and so quite naturally appear to the revolutionaries to be a ripe target. But grateful as I am for their constructive recommendation that I loosen my grip on dogma, I have already done a quite searching survey of the novel ideas they celebrate and found that, with a few fine exceptions, it is the revolutionaries who need that salutary icewater thrown in their faces. They typically combat a caricature of neo-Darwinism, and I find more dogmatic zeal in the neo-Lamarckian, “holistic” circles than in the calmly measured defenses of orthodoxy they have provoked. By my lights, the pre-eminent role of genes is secure, but this does not amount to the “gene centrism” they condemn, and yes, the Central Dogma is alive and well in biology, somewhat clarified and adjusted, but not abandoned, and for good reason. The Central Dogma declares that information flows one way only, from DNA to protein, never from protein back to DNA: an organism can “read” DNA as a recipe for a protein but cannot “read” protein as a recipe for DNA that makes it into the germ line. Francis Crick coined the term in 1956, and wrote in his autobiography (1988):

I called this idea the central dogma, for two reasons, I suspect. I had already used the obvious word hypothesis in the sequence hypothesis, and in addition I wanted to suggest that this new assumption was more central and more powerful. . . . As it turned out, the use of the word dogma caused almost more trouble than it was worth. Many years later Jacques Monod pointed out to me that I did not appear to understand the correct use of the word dogma, which is a belief *that cannot be doubted*. I did apprehend this in a vague sort of way but since I thought that *all* religious beliefs were without foundation, I used the word the way I myself thought about it, not as most of the world does, and simply applied it to a grand hypothesis that, however plausible, had little direct experimental support. (p. 109)

It has little *direct* experimental support, but it makes sense of many phenomena, and there are arguments (see, e.g., Dawkins, 1982, in *The Extended Phenotype*) to the effect that Lamarckian violations of it could not in the long run be beneficial. In other words, it seems to be a Good Trick. And no known counterexamples yet, in spite of many eager false alarms (as I discuss briefly below).

I have also already anticipated their advice to include other selective algorithms and processes in my purview, having been inspired, as they were, by Peter Godfrey-Smith's (2011) ideas about Darwinian Spaces and the "de-Darwinization" of natural phenomena. This comes as no surprise to them; they approvingly cite my review (2011) of Godfrey-Smith, correctly seeing it as a sign of "growing pluralism," but they will find much, much more in this direction in Dennett *BBB* (2017). In fact, one way of gauging the aptness of F & P's critique of my previous work is to note how many points in *BBB* are foreshadowed in their essay.

But there remain a few points of disagreement worth highlighting. In addition to the exposition of the immune system (as currently understood, with many gaps), they provide brief accounts of the posttranscriptional mechanisms, such as alternative splicing, frameshift, micro-RNA, and DNA methylation, and say that a major implication of these discoveries is that "the informational flow is not unidirectional (going from the DNA to its products) but is the result of many different mechanisms that use DNA sequences in different ways to respond in a flexible fashion to the environment." There is a crucial equivocation in their use of "unidirectional": the wildly branching developmental paths described ("going off in all directions" one might say) are not in themselves a violation of the unidirectionality declared in the Central Dogma. In support of their claim, F & P quote Griffiths and Stotz: "The significance of this is that these molecules relay specific difference making (instructive) environmental information to the genome." This sentence seems to be making something akin to a type-token error. To a first approximation, there is a copy—a token—of every gene in your genome in every human cell in your body. The specific individual RNA molecules being "read" or "interpreted" by the ribosomes or other developmental machinery yield different "readings" depending on what these "instructive" molecules impose on them on the occasion, and this leads to all manner of phenotypic plasticity in response to developmental variations. (See Haig, unpublished, for a lucid and eye-opening account.) It might seem that this could be interpreted as relaying environmental information *to those tokens*, but the tokens don't change, don't react; the reading of them does. (If I whisper "smut" in your ear as you read *Lady Chatterley's Lover*, it may well have a profound effect on your reading, but the text picks up nothing, and if it is passed to another reader will be the same text you read.) Among the effects wrought by a different reading of one token during development can be



effects on the reading of other tokens of genes during that developmental process. (Among the effects of your reading *Lady Chatterley's Lover* may be provoking others to read it, or to read other books by Lawrence; if your enthusiastic reaction triggers the creation of a best-seller your reading certainly “relays . . . environmental information” to the whole genome-reading process, but it has not (yet) been shown that these interactions have any effects on the tokens of the genes sequestered in the germ line, which is what matters for Lamarckianism. The tokens of all those genes in the sperm and eggs are not adjusted in any way by any incoming information. There can be hugely indirect effects, of course, on the fate of *later tokens* of those genes, mediated by all the impingements of the environment on the whole population of organisms in the gene pool—that’s how natural selection selects, by favoring the gene types whose tokens are most adaptively read by processes in all the organisms. (See also Haig, 2012, for a brilliant perspective on these issues).

When F & P next turn to apply the lesson to memetics, much the same problem affects their proposal: their chief spokesperson, Maurice Bloch, somewhat caricatures memetics in his critique. While Dawkins and I have described memes as individuable, countable units of some yet-to-be-fully-defined sort (cf. the original definition of genes, before DNA was on the scene), Bloch sees them (quoting F & P, not Bloch) as “discrete explicit and disincarnate representations, accessible directly through introspection to members of a culture.” (p. 274). Bloch, in contrast, “sees culture as formed by ‘lived-in models.” (p. 274) I agree that this rather Cartesian vision of culture Bloch describes won’t do, and in *BBB* I go to some lengths to show that memes can enter a vector unheralded and unsuspected, generate structures controlling psychological traits and deeds that are not at all word-like, and, indeed, proliferate by a variety of processes only some of which are strongly Darwinian. Words are, I argue, the best examples of memes precisely because they are so detectable, countable, distinguishable, and portable-without-major-adjustment. They are the “model organisms” of Memetics, easy to study now because they have been studied so much already. Sequoias and kangaroos are not much like fruit flies or *C. elegans*, model organisms for biology, but the same algorithms apply in ways harder to discern. And I agree with Bloch, and F & P, that there are different types of knowledge, not all propositional. In fact, *know-how*, or procedural knowledge, is in my opinion, a more basic form of knowledge than factual, propositional knowledge, and memes play as great (or a greater) role shaping know-how as in bestowing facts (and falsehoods).

So F & P should take satisfaction in the fact that their advice about “what Memetics-D should do” has already being taken to heart by me in most regards. It never hurts to be reminded of why what you’ve decided to do is a good move, and sometimes it brings up important new points.



## Works Cited

- Crick, F. (1988). *What mad pursuit: A personal view of scientific discovery*. Books in the Alfred P. Sloan Foundation Series. New York, NY: Basic Books.
- Dawkins, R. (1982). *The extended phenotype: The long reach of the gene*. Oxford, UK: Oxford University Press.
- Dennett, D. C. (2011). Homunculi rule: Reflections on *Darwinian populations and natural selection* by Peter Godfrey-Smith. *Biology and Philosophy*, 26, 475–488.
- Dennett, D. C. (2017). *From bacteria to Bach and back*. New York, NY: W. W. Norton.
- Haig, D. (2012). The strategic gene. *Biology and Philosophy*, 27, 461–479.
- Haig, D. (unpublished). Making sense: information interpreted as meaning. Retrieved from <http://philsci-archive.pitt.edu/id/eprint/13287>

# 10.1 PLANNING AND PREFIGURATIVE POLITICS

## THE NATURE OF FREEDOM AND THE POSSIBILITY OF CONTROL

Bryce Huebner

Like most animals, humans will learn to associate actions with the rewards and punishments that typically accompany them. By mining past experience for information that will improve future predictions, we track fluctuations in the distribution and value of rewards, monitor changes in the probability of gains and losses, and produce moment-to-moment estimates of risk and uncertainty (Montague, 2006). But while these backward-looking systems explain a surprising amount of human behavior, they make our ability to plan ahead and to exert forward-looking control over our behavior somewhat mysterious.<sup>1</sup> Yet, before visiting a new city, I often search for information about the best cafés, and I make plans to visit some. By focusing my attention on some cafés, and removing others from further consideration, I probably miss some hidden gems. But planning ahead also reduces the uncertainty inherent in finding good coffee in an unfamiliar city, and minimizes the risk of drinking bad coffee. When I travel with friends, we often plan together. This prevents forms of indecision and dispute that would prevent us from drinking coffee; and sometimes it allows us to uncover sources of excellent coffee that I would have missed.

Coffee is important. But such phenomena are the tip of a much larger iceberg. Many humans can imagine novel possibilities, decide which goals to pursue, and decide which means to employ in pursuing

---

1. In this chapter, I focus primarily on the possibility that explicit, propositionally articulated thought and speech can play a role in action guidance. I think that such forms of action guidance are rare, and difficult to sustain in the face of ongoing subpersonal processing and unconsciously computed expectations. But I do think that such states can facilitate forms of cognitive control, and I hope to offer some insight into how they are able to do so.

them; and as a result, we can improve our situation, instead of just acting to minimize the effects of our contingent learning histories (Dennett 1995, 377ff; 2003, pp. 266–268). The ability to look ahead gives us control over our actions, and it does so by opening up elbow room to act in accordance with our values and ideals.<sup>2</sup> My aim in this chapter is to clarify some of Daniel Dennett’s significant insights about these distinctively human forms of freedom. Specifically, I want to explore his claim that moral agency often depends on a little help from our friends (Dennett, 2003, p. 135). I argue that by deliberating together about collective actions, we can open up forms of prefigurative practice that change the available possibilities for action. But I build up to this claim slowly, and perhaps surprisingly, by “turning the knobs” on Michael Bratman’s (2014) theory of planning agency.

## 1. Intentions as Planning States

Bratman (2014) conceives of intentions as planning states, which play a critical role in our “internally organized temporally extended agency and to our associated abilities to achieve complex goals across time, especially given our cognitive limitations” (p. 27). Intentions terminate reasoning about which ends to pursue, prompt reasoning about the best means of pursuing them, and serve a coordinative role in temporally extended and collaborative actions. And as Bratman notes, intention-governed cognition typically unfolds in two phases: we first form a partially abstract plan, then we anchor it to a particular situation, specifying what must be done to achieve our ends, given our current circumstances. This makes our plans flexible yet binding, allowing them to structure our decisions and guide our ongoing behavior.

In forming intentions, we should be sensitive to information that might affect the feasibility and desirability of the actions we are considering. But this requires attending to norms of consistency and coherence: We aim to make our plans internally consistent; we aim to make them consistent with our beliefs about the world; and we aim to make them consistent with our other plans and projects. In forming an intention to teach a course on the evolution of language next spring, for example, I commit to avoiding conflicts with other courses and to updating my plan if I learn of new theoretical or scientific perspectives; this ensures that my plan remains consistent with available information. I also commit to revising or abandoning my plan if I learn that it is internally inconsistent (e.g., because

---

2. I would like to thank the members of a class Free Will, Intention, and Responsibility, at the Jessup Correctional Institute in Maryland, for helping me to see the power of this idea.

the issues require too much background knowledge for the students), unlikely to succeed (e.g., because students are uninterested), or in conflict with my broader network of plans and projects (e.g., if I will be unable to meet important writing deadlines if I try to develop a new class). Finally, forming this intention should lead me to stop thinking about what to teach and to start thinking about the best means of teaching. And by settling on plausible means for executing my plan, I provide myself with the time to do the necessary reading, and the time to construct a coherent course narrative.

In general, forward-looking plans evoke present-directed intentions, which motivate us to pursue our ends and to continue to do so until we reach them. They shift our attention to the guidance of behavior, in ways that allow us to maximize efficiency and effectiveness (Pacherie, 2006, p. 150). And in this respect, they impose a cognitive filter on thought and behavior, highlighting action-relevant information and action-relevant facts about our current situation; and as a result, we pursue our intended actions unless we learn that they will have problematic collateral effects. Having formed an intention to drink some coffee when I finish writing this paragraph, I will typically take action to do so. But I will abort my plan if I realize that the only coffee around is Ruth's soy cortado; though my current informational state is goal directed, I remain sensitive to inconsistencies with other goals (e.g., not making my partner angry). Likewise, I will abandon my plan to teach particular articles if I learn that students don't have the background to understand pushdown architectures. But in general, unless we run into trouble, our intentions guide behavior in accordance with our forward-looking plans because they represent situations we want to bring about, as well as paths to achieving the goals that we have committed to (Bratman, 2008, p. 53).

A wide range of psychological data confirm that we think about things differently when we deliberate about which ends to pursue and when we are trying to find the best means of implementing our plans and organizing our ongoing behavior (see Gollwitzer, 2012, for a review). Deciding which ends to pursue tends to evoke a *deliberative mindset*, in which attention becomes focused on information about the feasibility and desirability of different options. In this mindset, we remain more open to alternative possibilities, more realistic about the prospects of success, and more sensitive to the pros and cons of different options. People in deliberative mindsets also tend to be more accurate in their assessments of how much control they have over future actions and outcomes, including the risk of being in a car accident, the risk of becoming depressed, the risk of developing a drinking problem, and the risk of being mugged (Taylor & Gollwitzer, 1995). In part, this is because deliberative mindsets evoke heightened receptivity to potentially relevant information, leading us to look for different causal explanations

of the things that happen in the world (Gollwitzer, 2012, p. 537). This makes sense. A person who is trying to decide what to do cannot know in advance where the decision will take them; so being receptive to information is an appropriate and functional task solution (Gollwitzer, 2012, p. 528). By being careful not to dismiss information that might be useful later, people in a deliberative mindset can thus become more sensitive to factors that affect the coherence of their ends, and somewhat more accurate in their estimations of how likely they are to reach their goals.

Once people settle on a goal, they tend to focus on goal-relevant information while ignoring or downplaying goal-irrelevant information (Gollwitzer, 2012, p. 534). In an *implementation mindset*, it makes sense to focus on the best means of achieving our goals; but this leads people to be more optimistic about their chances of success, and more partial in their judgments about the desirability of the ends they pursue. People in such mindsets tend to overestimate the control they have over various outcomes, and underestimate their vulnerability to various risks (Taylor & Gollwitzer, 1995). By focusing on information related to the achievement of their ends and ignoring other sources of information, we can more rapidly determine the best way to achieve our goals, and shield ourselves from the distractions imposed by competing considerations (Gollwitzer, 2012, p. 528). As a result, we become more resolute in the pursuit of our goals—especially where the perceived feasibility of a task is low, but the desirability is high, or vice versa (Brandstätter & Frank, 2002). But this comes at a cost. We can easily miss important information if it is not within the range of information to which we are currently attentive.

## 2. Foreknowledge and Freedom

Dennett (2003, p. 251) often argues that one of the most distinctive facts about human agency is that we can make requests of ourselves, and at least sometimes we comply with them. This requires more than just acting in accordance with a command. My cat might jump off the table when I tell her to get away from my coffee, but she doesn't understand the sounds I produce. And when she ignores those sounds, she is not declining my requests. Furthermore, she cannot make requests of herself. She cannot decide to limit her food consumption to improve her jumps; and more generally, she cannot make explicit plans and use them to guide behavior. By contrast, when we intend to do something, our plans place normative constraints on behavior, and deviations from our plans call for reasoned explanation (Bratman, 2008). Like failures to comply with accepted requests, the failure to follow through on our commitments feels bad, and it calls for a justification.

Of course, failures to follow through on our plans can arise in many different ways, and each calls for a different response. Follow through is difficult when we are tired, busy, or overwhelmed; and even minor stress can lead us to abandon computationally taxing forms of explicit planning in favor of habitual and Pavlovian forms of behavior (Crockett, 2013; Schwabe & Wolf, 2013). But in such cases, our failures are mechanical, and a plausible explanation can proceed from within the design stance. We can also change our minds before we act, even when we have settled on a plan. My plan to make a cup of coffee, for example, may fail to trigger a present-directed intention when I decide that I don't have time to do so or when I decide that a sixth cup isn't necessary. But there are many other reasons why we don't follow through on our plans, including picking tasks that were too intellectually or physically difficult to complete, or too conceptually or physically distant from our current situation. These are paradigmatically rational decisions, and I can explain and justify them, both to myself and to others. But sometimes I just lack the "willpower to get started, to stay on track, to select instrumental means, and to act effectively" (Gollwitzer, 2014, p. 305). These failings are agential but not rationally justifiable, and our ability to control them reveals an important sense in which foreknowledge can yield control over our ongoing behavior.

By looking ahead, and imposing forward-looking constraints on behavior, we can sometimes find better ways to act in accordance with our goals and values. Precommitting to an action can decrease the likelihood of pursuing immediate rewards, at a cost to our future interests (Ainslie, 2001), and it can decrease the likelihood of cheating or shirking when the going gets tough (Schelling, 1966). As Dennett (2003) notes, precommitment raises the stakes and "changes the task of self-control we confront" (p. 207). Precommitting to an action shields us against temptations that might lead us to abandon our chosen ends (Crockett et al., 2013). And since intentions are conduct-controlling pro-attitudes that we are inclined to retain without reconsideration, they should be able help us solve more complex intrapersonal commitment problems (Bratman, 1987, p. 20). And they do.<sup>3</sup>

Precommitting to a plan delegates control over the initiation of an action to a situational cue, creating a strong associative representation that links that cue to a relevant response (Gollwitzer, 2012, p. 537). By specifying a precise, situation-specific action-plan, we automate the translation of future-directed intentions into present-directed intentions (Adriaanse, Gollwitzer, De Ridder,

---

3. As John Gavazzi (personal communication, September 15, 2015) reminds me, precommitment strategies also play a prominent role in clinical environments, where people are trying to change specific behaviors such as overeating or smoking.

de Wit, Kroese, 2011; Crockett et al., 2013; Gollwitzer, 1999). The method for doing this is surprisingly simple. Rehearsing “implementation intentions,” simple if-then plans that specify precise trigger cues and behavioral responses, instigates a form of reflexive action control that shields behavior from temptation and distraction, and counteracts the effects of habitual behavior patterns (Gilbert, Gollwitzer, Cohen, Burgess, & Oettingen, 2009). I can increase the likelihood of taking my multivitamin each morning, for example, by forming the following plan: “When I have my morning coffee, I will take my multivitamin.” While I might lack the processing power to remember to take my multivitamin in the morning, this intention creates a significant association between a particular situation and a particular action, making it easier to remember to act, by increasing the salience of my intention in the relevant context (Gollwitzer, 2014). Similar intentions can affect the ways that I think, my affective responses, and the behavioral dispositions I adopt; they can also reduce disruptive influences that originate from innate action tendencies, learned habitual responses, behavior priming, or entrenched social prejudice (Gollwitzer, 2014, p. 311). By imagining a future in which we act in a particular way, we can recruit associative mechanisms to guide our behavior in accordance with the plans we form.

Similar effects arise in many cases, and a large meta-analysis has revealed substantial effects of implementation intentions on goal attainment across domains as diverse as consumer decisions; academic achievement; health-relevant behavior; and morally significant phenomena, such as being more egalitarian and more pro-social ( $d = .61$ ; Gollwitzer & Sheeran, 2006). And more targeted meta-analyses have confirmed these effects for behavior relevant to healthy eating and physical activity (Adriaanse, Vinkers, De Ridder, Hox, & de Wit, 2011; Belanger-Gravel, 2013). Most surprisingly, implementation intentions can moderate the perceptual tendency to mistake a cell phone for a weapon in a racialized context (Mendoza, Gollwitzer, & Amodio, 2010; Webb, Sheeran, & Pepper, 2012). By rehearsing the plan “if I see a black person, I will think ‘safe,’” people can bring their reflexive behavior into line with antecedently held egalitarian goals. But as powerful as they are, implementation intentions only guide behavior when we already have strong commitments to particular goals. Since they operate by linking high-level plans to target cues, they can only yield present-directed intentions when we are disposed to prefer the actions we are attempting to control. Put differently, implementation intentions allow us to maintain top-down control over behavior by using foreknowledge of future contingencies to arrange associative thought in ways that prevent us from pursuing worse options though we know better ones are available (cf., Spinoza, 1677/2002). Consequently, the strongest effects of implementation intentions arise

when people use cues they already consider critical to an action, and link these cues to a behavioral response they already believe to be relevant to the achievement of that goal (Gollwitzer, 2012, p. 541).<sup>4</sup>

### 3. Planning and Cognitive Architecture

Neuroscientists have long known that even highly routinized tasks, such as picking up a coffee mug, require integrating goal-based representations, perceptual inputs, and ongoing proprioceptive and evaluative feedback to yield representations encoded for use by motor systems, control loops, and emulator circuits (Akins, 1996, p. 354; Mahon & Caramazza, 2008). As Elisabeth Pacherie (2006, 2013) argues, this fact suggests an important role for motor intentions in the guidance of behavior. Motor mechanisms don't require the guidance of explicit plans; they operate reflexively by minimizing discrepancies between predicted and actual motor states and making micro-adjustments whenever such discrepancies arise. But just as importantly, these systems are embedded in a computational hierarchy and integrated with systems that constrain their behavior in some cases. At the top of this hierarchy, forward-looking intentions, structured as explicit plans, lead us to pursue goals that we have settled upon rationally; at the mid-level, present-directed intentions arise as we flesh out situation-relevant specifications of those plans—guiding behavior in accordance with our forward-directed intentions; and at the bottom, motor intentions initiate and guide behavior in real time, facilitating moment-to-moment control over our ongoing action.

There is a rapidly expanding consensus that the human brain is a hierarchically organized predictive machine, which consists of numerous error-driven learning systems that each have their own tasks, and their own models of those tasks (Dennett, 2015; Friston, 2009; Howhy, 2013; Rescorla, 1988). Systems closer to the bottom of this hierarchy traffic in more precise sources of information, and processing becomes more abstract and more conceptual toward the top of the hierarchy. But although each system provides input to the system above it, the goal of cognitive processing is not to build a more useful representation of the world as information is propagated upward. Instead, each system attempts to predict the inputs it will receive, given its model of the world. These predictions flow

---

4. While implementation intentions only succeed where a goal is perceived as important, and preferable to the existing alternatives, there are ways of affecting these states as well. By imagining a desired future and reflecting on the facts that stand in the way of reaching that future, people can enhance the cognitive relevance of the desirable features of an action, and in this way, they can highlight the feasibility of their plan (Oettingen & Gollwitzer, 2010). I return to this point in the section, "Help From Our Friends."



downward, resolving ambiguous data without additional search. At the same time, each system generates error signals when surprising data are encountered; these error signals flow upward, recruiting new top-down predictions to better accommodate the incoming data. Over time, this bidirectional flow of information allows the brain to search for “the linked set of hypotheses that, given the current raw sensory data, make that data most probable” (Clark, chapter 7.1, this volume).

Andy Clark has shown that the predictive coding framework can clarify the extent to which top-down predictions affect perceptual experiences. As we move about the world, we extract and complete perceptual patterns in accordance with our perceptual expectations. And often, these expectations allow us to act even though the incoming perceptual signal is noisy and even though there isn’t enough incoming information to guide situation-relevant action. Lisa Feldman Barrett (2014) has argued that top-down expectations also have a significant impact on the emotions we experience, as well as on the action tendencies that are recruited as we move through the world. And I’ve argued elsewhere that top-down expectations often impact our ability to conceptualize possibilities for socially significant action (Huebner, 2016). Here, I focus on expectations that take the form of planning states.

A highly salient example of this occurs when expectations about violent Black males and the likelihood of violent crimes in particular neighborhoods guide action-oriented perception. A person passing through a neighborhood that evokes these expectations will experience increased awareness of potential threats, negative affective valence, and enhanced accessibility of stored associations between race and violence; this will often trigger the construction of action plans designed to prepare this person to navigate potential threats (Wilson-Mendenhall et al., 2011), and it will greatly increase the likelihood that this person will *see* a cell phone as a gun (Barrett, 2015). In this situation, ambiguous visual stimuli (e.g., a barely glimpsed cell phone) are more likely to be resolved in accordance with racialized expectations, and thus will feed back into action-oriented processing, thereby increasing the likelihood of acting as though the innocuous object is a threat. This claim might seem surprising if we assume that incoming data fully determine what we see and what we are motivated to do. But conceptualizing the brain as a behavioral guidance system, which is designed to use ambiguous data to navigate a dangerous and dynamic world, predicts that higher-level expectations will often be used to resolve perceptual ambiguities in favor of expectations about what is likely to happen next; where we expect a pattern, the brain will try to complete it. And sometimes the results are awful!

But what role do planning states play in this kind of processing? As I noted earlier, implementation intentions can have an impact on everything from

motivation to ways of thinking and affective responses, in ways that yield robust behavioral dispositions—and they can do so even when perceptual cues are seen briefly or tracked subconsciously (Gollwitzer, 2014, p. 308). In forming such intentions, we generate new top-level expectations, which explicitly link a target cue with a relevant response. These expectations have conceptual structure, but they also provide a top-down signal that can guide the response of lower-level systems to ambiguous sensory inputs. Implementation intentions serve as models, which can be fed downward to lower-level systems (Huebner, 2016; Kim et al., 2011), and so long as error signals that lower-level systems compute do not routinely violate these expectations, these models, which are constructed at the top level of the Tower of Generate and Test, will continue to constitute part of the overall hypothesis that best approximates the incoming data. Over time, unless error signals require adopting alternative hypotheses, cognition should converge on a linked set of hypotheses that make actions cued by implementation intentions the most plausible. But as George Ainslie (2001) suggests, failures to act in accordance with our plans should be disastrous, as they will cause this linked set of hypotheses to collapse; and where the world routinely provides perceptual information that contradicts or expectations, they should shift to become consistent with the world, even if they are less consistent with our top-level goals. So our intentions sit atop the Tower of Generate and Test, serving as top-level expectations that can guide behavior when doing so is necessary.

In light of these claims, I maintain that Dennett is right about several things: multiple processes, operating at multiple levels, guide ongoing behavior; planning depends on explicit representations that lie at the peak of the Tower of Generate and Test; explicit representations only need to arise at the top level of the Tower of Generate and Test; and plans allow us to enact forms of forward-looking control by using socially situated conceptual knowledge to reshape our responses to the world from the top-down. By binding our understanding of the world to particular situations and activities, we can rely on perceptual, evaluative, and motor representations to guide our moment-to-moment behavior; but we also rely on “revisable commitments to policies and attitudes that will shape responses that must be delivered too swiftly to be reflectively considered in the heat of action” (Dennett, 2003, p. 239). And I thus contend that we should accept something like the following view of human agency (though the details may shift as we learn more about the hierarchical structures that are operative in the guidance of goal-directed behavior):

- Future-directed intentions, structured as planning states, yield high-level expectations and allow us to impose goal-directed structure on our actions.

Our future-directed intentions typically evoke present-directed intentions because they provide top-down pressure on the computational cascade that triggers goal-directed behavior.

- Absent-minded and *akratic* behaviors arise when we default to habitual or innate responses patterns, and when encounters with the world lead us to pursue worse options though we know there are better ones available (Spinoza, 1677/2002). But with foreknowledge of the conditions under which these failures of agency are likely to occur, we can use top-level intentions to prevent such failures (using implementation intentions and precommitment strategies).
- Present-directed intentions can also arise at the mid-level of computational processing, even without future-directed intentions to guide them. This will produce spontaneous intentional behavior or rational forms of habitual behavior (Bratman, 1987, p. 119ff; Tollefsen, 2014). Sometimes free-floating rationales, which are not represented explicitly, guide such actions; and sometimes internalized rational expectations that allow us to navigate our social world guide such actions habitually (as we'll see, this yield problems for us as agents).
- Finally, motor intentions can sometimes be produced without evoking present-directed intentions, and this will yield forms of goal-consistent behavior without rational control. Such behavior should be experienced as reflexive and automatic, but post-hoc rationalization may nonetheless allow us to treat such behavior as resulting from existing plans, thereby yielding illusions of conscious control (Pacherie, 2006).

Unfortunately, I worry that the kind of freedom that foreknowledge provides on this picture isn't quite as robust as we might hope. And I am less sanguine than Dennett (2003) about the claim that we are the "authors and executors of these policies, even though they are compiled from parts we can only indirectly monitor and control" (p. 239).

#### 4. A Killjoy Interlude

While implementation intentions provide evidence of distinctively human forms of action guidance (Holton, 2009), the reason for their success also reveals their most troubling limitation. Expectations can significantly affect perception and action, but they aren't created out of whole cloth. We acquire particular expectations though our encounters with the world; and we evaluate and select actions on the basis of expected rewards, and evaluative facts about our current state and

current motivations (Dehaene & Changeux, 2000; Montague, 2006; Polania, Moisa, Opitz, Grueschow, & Ruff, 2015). But we live in a world that is thick with structural racism, sexism, ableism, trans\*phobia, and xenophobia. So as we watch TV and films, read novels and blogs, and walk through familiar and unfamiliar spaces, we are bombarded with statistical “evidence” that fosters the construction of exclusionary assumptions; far more troublingly, it’s rare for most of us to encounter situations that would recruit and sustain the anti-kyriarchical expectations that many of us hope to formulate.

These social structures impact our highest-level values and ideals, both the ones that guide our behavior unconsciously, and the ones that we consciously avow. We see the latter when philosophers express preferences for particular questions and methodologies, treat certain things as signals and others as noise, and assume that answers must take a particular form. As Dennett has long argued, people often acquire these attitudes as part of their training as philosophers. But the problem runs much deeper. As Nathaniel Coleman (2015) argues, the discipline of philosophy has been *whitewashed*. The institutional structures that govern academia foster expectations grounded on racialized assumptions that are rarely acknowledged, and less often challenged. Similar problems arise in most areas of our cognitive and social lives. Because we attune to social practices, our high-level attitudes and low-level reactions entrain to statistically prevalent norms and practices. This leads to stable institutional structures against which future attitudes and behavior can attune; and as a result, judgments about which plans are feasible and desirable tend to be structured around cognitive biases that have become entrenched in social norms and practices that guide behavior in ways that are beyond cognitive control.

Even worse, attempts to transcend our biases by constructing expectations that run contrary to dominant social norms typically give way to actions that accord with the norms and practices we are trying to overcome. This happens as actions driven by countersocial expectations are met with feedback suggesting that we are making mistakes (Klucharev, Hytonen, Rijpkema, Smidts, & Fernandez, 2009; Klucharev, Munneke, Smidts, & Fernandez, 2011); and when error signals continually arise, new hypotheses are recruited to make our behavior more consistent with the world we inhabit. Put much too simply, as the brain searches for the linked set of hypotheses that make incoming data most plausible, our expectations will shift toward statistically common patterns that we hope to overcome. Our degrees of freedom are therefore socially limited, and the freedom we have to plan ahead is constrained by the norms that govern our social lives. Put much too bluntly, we are free to conform in the long run, even though we can resist in the short run.

As far as I can tell, this is an implication of Dennett's (2003) hypothesis that a "proper human self is the largely unwitting creation of an interpersonal design process in which we encourage small children to become communicators and, in particular, to join our practice of asking for and giving reasons, and then reasoning about what to do and why" (p. 273). This is not to deny his claim that we are "capable of altering course at any point, abandoning goals, switching allegiances, forming cabals and then betraying them, and so forth" (p. 155). But we can only do so on the basis of goals and values that we have acquired through our interactions with the world. And where problematic patterns are pervasive, and where we act in ways that feel right because they are statistically regular, we will only hit upon normatively valuable practices by accident or luck.<sup>5</sup>

Nonetheless, there is something right about Dennett's claim that we have more degrees of freedom than other animals do. We can get stuck on the local peaks of an adaptive landscape, and we rarely consider the possibility of better options. But we do have the capacity to imagine options that aren't currently available (Dennett, 2003, p. 267). As I see it, the problem we face is a Darwinian problem: it is hard to sustain novelty. I have argued elsewhere that doing so requires building a world around preferable values and ideals (Huebner, 2016). But if the arguments I have advanced here are roughly right, judgments about what kind of world is preferable will also be governed by the social world we inhabit. Nonetheless, I believe that *planning together* can sometimes open up new possibilities. By working together, we can imagine another world, and attune to local ways of thinking that can prevent the kinds of cognitive backsliding to which Bayesian thinkers often fall prey. My aim in the remainder of this chapter is to explain how this is possible.

## 5. Using Cognitive Prosthetics

Let's start with a banal case of plan-driven behavior: making an excellent cup of coffee with a Hario v60. This plan is complex, and to successfully execute it I must perform the following steps: place a v60 atop a mug, place a filter inside the mug, boil water, rinse the filter, discard the rinse water, grind the proper quantity of beans and place them in the filter, make a divot in the center of the grounds, pour an ounce of near-boiling water over the grounds and let them bloom for 30 seconds; and then pour water over the grounds, in concentric circles, until the mug is filled with delicious coffee. Some of these steps are more optional than others. I can decide not to rinse the filter, make a divot, or let the grounds bloom (though

---

5. I take this to be one of Plato's insights in *Gorgias*, and the core of Spinoza's worries regarding human bondage.

these choices will affect the quality of my coffee). And I can skip these steps without losing track of which steps must still be executed, even where they have always been carried out in the past (Fitch & Martins, 2014). Furthermore, I can boil the water or grind the beans first, with little impact on the remaining steps. These steps are required, but their order is somewhat optional; they must only occur prior to the bloom. But I cannot pour the water unless I have placed the filter and beans in the v60; like many actions, this one is hierarchically structured, and some actions must be carried out before others. Finally, I can pause after completing any step prior to the bloom, and resume without compromising success. These facts seem to suggest that I am able to mark where I am in the process, and track the dependencies between various subtasks; and this seems to require complex internal capacities for storing and manipulating internal representations (perhaps using something like a pushdown stack architecture; cf., Fitch & Martins, 2014, p. 96).<sup>6</sup>

But before we get too comfortable with claims about such forms of internal guidance, we must note that we often supplement our representational capacities with cognitive prosthetics; and our plans become sparser as the information in our environment becomes more richly structured (Simon, 1969). As Zoe Drayson and Andy Clark (forthcoming) argue, people with Alzheimer's often rely on this fact to compensate for internal memory deficits. They enrich their material and social environment by placing notes around their houses specifying what to do and when; they label things; they set up reminders to organize their behavior; and they rely on loved ones to help navigate their lives. This social and material scaffolding also helps them to make plans and to find ways of living more autonomously. But this is also the standard situation for neurotypical people; and forms of social and material scaffolding play a significant role in our ability to remember the past, plan for the future, and consider alternative possibilities (Kosslyn, 2006).<sup>7</sup>

---

6. Intriguingly, sequential planning can also be compromised, while the ability to carry out habitual actions, and even component actions, is preserved. In rare cases, *action disorganization syndrome* can lead a person to omit tasks, do them in the wrong order, or perseverate on a single task (Humphreys, Forde, & Riddoch, 2001; Jackendoff, 2007). They might forget to put the filter in the v60, pour water directly into the mug, or continue to pour water even when the cup is full. But while forward-looking plans do seem to be realized by systems beyond those employed required for habitual learning (Lashley, 1951), there is an ongoing debate over whether planning requires systems that construct hierarchical representations (Fitch & Martins, 2014; Rosenbaum, Cohen, Jax, Weiss, & Van Der Wel, 2007), or whether Bayesian systems dedicated to probabilistic inference suffice (Botvinick & Toussaint, 2012). At this point, the underlying architecture remains unclear. Thanks are due to Joey Jebari for pushing me to clarify this issue.

7. An initial attempt at understanding the kinds of neural processing that make cognitive offloading possible has recently been carried out by Julia Landsiedel and Sam Gilbert (2015). They found increased BOLD response in a network of "task-positive" regions when people

Consider Barika's use of a list of processes, written on a sheet of paper, to guide her coffee-making behavior. At each step, she must check the sheet, and carry out the next step. This makes her action rigid and inflexible, but the structure of her informational environment minimizes the information she must remember. Depending on her working memory capacities, and her ability to think clearly in morning, this may be the best way to successfully make coffee. She still needs to track where she is in the task, and she still needs to flesh out details of her plan to suit her current situation. But by planning to have this list of processes on hand, Barika can increase her degrees of freedom by changing the world instead of changing herself. Of course, changes to the self will often happen downstream. When I acquired a *rakweh* in Lebanon, I had to search the Internet for instructions, and follow them rigidly until I could make coffee with this simple device. Over the course of many weeks, I began to tweak the plan I had read about, and found ways to manipulate the component actions to develop a more flexible skill for making excellent Lebanese coffee (both with and without cardamom). But the initial planning phase depended on externally stored representations. And in many cases the representations we exploit remain outside the boundaries of skin and skull. Many of us now exploit the Internet as a cognitive prosthetic for developing forward-looking plans, and guiding online behavior; as a result, we tend to think about computers when we are presented with hard questions, and we are less likely to encode information if we believe it will be available on-line (Sparrow, Liu, & Wegner, 2011). This, too, can open up degrees of freedom that would otherwise be unavailable.

More interestingly, we can rely on other people to scaffold our ongoing behavior; and collaboration can also expand our degrees of freedom. Suppose two people are trying to navigate an unfamiliar city, in the dark, using a small map they have received at a conference (Gross, 2012). Zoe struggles with spatial reasoning, and she can't quite figure out the relationship between the map and the streets; Phred has a hard time reading, and she has difficulties making out the words that are written on the map. But Zoe and Phred also have complementary capacities. Zoe can read the map as well as the street signs, and Phred can understand the spatial relations on the map and their relationship to the

---

remembered delayed intentions, as well decreased response in the so-called default network. But when participants set external reminders for themselves, the deactivation of the default network was strongly reduced where there was a larger memory load, even though there was no parallel reduction in task-positive activation. They suggest that this is because medial rostral prefrontal cortex (PFC) activity is associated with externally cued rather than self-initiated activity. For our purposes, what is most interesting in these data is the suggestion that internal and external sources of information are being integrated to guide online behavior.



streets they are walking along. By talking to one another, and developing meshing subplans to guide cooperative behavior, they can successfully navigate the city *together*.<sup>8</sup>

Bratman (2014, p. 32) argues that joint intentions tend to arise in collaborative groups in which people (a) intend to act together (b) on the basis of beliefs about what the others intend, (c) when their meshing subplans are common knowledge, and (d) when they are mutually responsive to one another's plans and actions. In this case, Zoe and Phred might both intend to get back to their hotel, with the other, in a way that's consistent with their distinctive plans and abilities. Each of them might believe that they will get to the hotel, as long as the other follows through on their plans; and as they walk, their joint intention helps them respond to what the other says and does, and prepares them to update their plans if anything goes awry. If so, their joint intention will allow them to achieve ends that neither could achieve easily on her own. Bratman (2014, p. 42) also argues that intending to act together triggers an inferential chain, which leads us to form intentions to play particular roles in bringing about joint-activities. For example, if Zoe and Phred make a plan to get coffee at 15:00 hours, Zoe will tend to form the intention to drink coffee with Phred at 15:00 hours, and this will trigger actions directed toward bringing it about that they drink coffee together. As they plan, she should be sensitive to norms of *interpersonal consistency*, which will focus her attention on considerations that affect the feasibility and desirability of their shared end (Bratman, 2014, p. 29). And as they act, Zoe's focus should shift to the role she plays in bringing about their joint-activity (Bratman, 2014, p. 64). Where things go well, this should also lead Zoe to provide assistance to Phred where doing so is necessary; and it should lead her to make adjustments to her plan in light of Phred's behavior. Bratman (2014, p. 89) argues that these forms of social responsiveness result from facts about Zoe's plans, given their social content, and given the demands of interpersonal consistency these plans entail. And this is consistent with the model of precommitment and planning that I discussed earlier; intending that *we act together* generates a high-level expectation, which can guide individual behavior in ways that constitute collective action.

---

8. The claim that agency and autonomy depend on or relationships with others and with institutional structures of social power and the forms of self-understanding that they engender is far from novel (Christman, 2004; Holroyd, 2011; Kukla, 2005; MacKenzie, 2014; MacKenzie & Stoljar, 2000; McLeod, 2002). But these issues have yet to take hold in discussions of the cognitive science of agency. In future work, I hope to flesh out the points of contact between accounts of relational autonomy and the cognitive science of agency—for now, this must remain a promissory note.



## 6. Acting Together

We must proceed carefully in thinking about the value of acting together. After all, statistical learning mechanisms often suffice to calibrate behavior against the structure of our environment, and to prioritize competing sources of sensorimotor information (Dehaene & Changeux, 2010; Shea, Boldt, Bang, Yeung, Heyes, & Frith, 2014). This is even true where multiple agents are making decisions in parallel. Imagine a busy cafe with overlapping workspaces. Without compromising the flow of high-quality coffee, Matthew might need to put a cup in a space that Terry was keeping clear, and Rachel might need to get some beans from a drawer behind Matthew. These baristas must continually update their motor intentions and adjust their present-directed intentions to accord with their interactive context. And to do so, each must track their own actions, the actions of other baristas, and the unfolding of their joint actions; and each of them must dynamically update their behavior and intentions in light of changes in this situation (Tollefsen, 2014, p. 15). But if baristas had to track all of this information explicitly, real-time coordination would be impossible (Kirch, 2006); and, this point generalizes, the flow of information in many real-world environments is often too rapid and too sparse to make explicit planning impossible (Blomberg, 2015; Tollefsen, 2005).

Fortunately, Deborah Tollefsen and Rick Dale (2012) have shown how real-time coordination can be sustained by subpersonal mechanisms that track multiple sources of information in parallel to bring the behavior of interacting agents into alignment. We often treat our social environment as a resource to be probed with epistemically directed actions, and we adjust our motor states against the stream of data evoked by these probes (Friston, 2009, p. 295; Kirch & Maglio, 1994). In interactive contexts, the behavior of others partially constitutes this environment, and we can probe others, track forms of behavioral feedback (including facial expressions, body postures, gestures, and more), and adjust our own behavior reflexively and automatically (Tollefsen & Dale, 2012, p. 392). Interpersonal epistemic action thus allows us to formulate present-directed intentions that allow us to act together, fluidly and dynamically, without explicit plans or shared representations (Tollefsen, 2014, p. 21; Wiltzky, 2015). Where individual differences in the construal of a problem or a situation are irrelevant, these present-directed intentions can arise as mid-level representations, guided by free-floating rationales that are not explicitly represented. This is probably the normal context in which human behavior unfolds, as expectations are rapidly and continuously updated to ready us for situation-relevant forms of thought and action (Barrett, 2014). But, social interactions recruit social expectations; and sometimes, as we conceptualize ourselves as participants in joint activities,

forms of joint attention and explicit planning are recruited to navigate social situations.

As our behavior aligns with the behavior of social partners, we often exhibit more interactive forms of joint agency, and this triggers “more expressions of working together, feelings of solidarity, and so on” (Tollefsen & Dale, 2012, p. 402). Over time, this can bring our explicit attitudes into alignment and make us more sensitive to the options available to us as a group; as a result, we may be led to think of things that are less likely to occur to observers of a collaborative activity. Acting together can focus attention on the intentions, reasons, and emotions of others, allowing us to rely on different sources of information in the construction of explicit plans (Gallotti & Frith, 2013, p. 162). To reason and plan together, people must “survey and convey to others their own thoughts, feelings, memories, and imaginings. This requires not only representational capacities, for which consciousness is surely not needed, but meta-representations that we can make publicly available—including embedded if–then meta-representations of what one would think, feel, or seek to do under hypothetical circumstances” (Seligman, Railton, Baumeister, & Sripada, 2013, p. 130). And as Dennett argues, the use of explicitly represented mental states can open up novel possibilities for action control and action guidance. We are now in a position to see how this occurs (cf., Sie, 2014, for a set of broadly similar suggestions).

The production of explicit representations satisfies a supra-personal cognitive demand: explicitly represented mental states can be broadcast to others, and this allows us to bring action-guiding systems into alignment across differences in learning histories and differences in the construal of the current situation; by broadcasting explicit representations, we can triangulate cooperative behavior across notional worlds (Shea et al., 2014). As Bayesian learners, we constantly attempt to generate the best model of the world we inhabit; in the process, we attune to the statistical contingencies we have encountered. Each interaction with the world shapes our experiences and expectations. And since we encounter subtly different aspects of the world, in different affective states, with different expectations, the models we construct are likely to diverge in multiple ways that will go unnoticed without careful heterophenomenological investigations. The problem is compounded by the fact that biological differences in impulsivity, risk aversion, reward sensitivity, and perceptual acuity will have an impact on the features of the world that we each attend to, yielding complex differences in our understandings of what is possible and what is impossible in the world that we all inhabit. So we must “talk” to one another to coordinate across differences in notional worlds, as well as differences in moods, expectations, and biases; but our success in doing so will often vary as a result of our openness to others, and our willingness to reconsider our initial construal of a situation.

I don't mean to oversell this point, as many differences between our notional worlds are irrelevant to individual and joint activities. Nonhuman animals can act together without relying on explicit representations (Couzin, 2009), and so can children, who do not possess robust capacities for mutual understanding (Tollefsen, 2005); strangers can push cars out of the snow together without planning to do so; and protests can emerge spontaneously, among people who only have weak expectations about what others will do (Kutz, 2000). Differences between notional worlds are also minor enough in many cases that they have no noticeable effect on attempts to navigate a shared world. We all live in the same world, and we attune to many similar reward contingencies and statistical regularities. This can create notional worlds that replicate hegemonic ideologies, and that shape evaluative judgments to accord with dominant power structures (Huebner, 2016).<sup>9</sup> But occasionally, differences in notional worlds do make a difference, and explicit representations must be recruited to understand someone who experiences the world differently from us. This happens when we interact with someone with different cultural assumptions, a different political ideology, or different patterns of engagement with a lived environment. And it happens where we want to collaborate, in a complex decision-space where we are unsure whether our collaborator views the current situation in the same way we do (Shea et al., 2014, p. 188). By making representations of our states available to others (using gestural or verbal reports) we can create publicly accessible signals that can be evaluated for significance relative to joint-activities from multiple perspectives.

Shared representations can impose structure on the world that wasn't there previously, by "freezing" contents to serve as conceptual anchors in a sea of dynamic thought (Clark, 1996). And once they are present in collaborative contexts, explicit representations can be used synchronically to coordinate ongoing group behavior, or diachronically to facilitate the revision and adjustment of previously stored or expressed representations (Shea et al., 2014). As a result of these forms of supra-personal cognitive control, groups whose members communicate are routinely able to outperform groups whose members do not, in cases as diverse as mock-jury deliberations, perceptual and motor tasks, and tennis games

---

9. Philosophers should pay more attention to the complex interactions between the material structure of our world, the structure of notional worlds, and the capacity to act freely. These discussions have been at the center of research on disability; they were essential to Frantz Fanon's discussion of colonial power and colonized psychologies; and they were the foundation of W. E. B. Du Bois's discussion of double-consciousness. A recent paper by Olúfẹ́mi Táíwò (in prep) helps to clarify how these factors affect rational control and agency; and I hope to pursue these issues in my next book.

(Shea et al., 2014, p. 189). More intriguingly, these forms of mutual responsiveness can help groups realize things that individual members would have missed, and they can increase the likelihood of revising or abandoning failed projects.

## 7. Sharing Implementation Intentions

Imagine a well-functioning academic department that is making a hiring decision. Each faculty member knows different things about the candidates, and each has their own opinion about who to hire. But they also share the goal of hiring the best candidate possible, and they agree on a set of conditions they would like to see satisfied. This group should make a better decision than any individual would make on their own, so long as they share and use the information distributed throughout the group. But groups routinely fail to capitalize on their informational advantages; they forget to share crucial information, and when they do share it, it's often ignored unless most group members already know about it (Wieber, Thürmer, & Gollwitzer, 2012, p. 278).<sup>10</sup> Fortunately, but groups can exploit the same action-guiding techniques as individuals, for example, when their members identify with a shared goal. Where the best alternative is difficult for individuals to see, but identifiable in light of a group's total information, forming an implementation intention ("When we are about to make the final decision, then we will go over the advantages of the non-preferred alternatives again") can increase the likelihood of choosing the best alternative (Thürmer, Wieber, & Gollwitzer, 2014). This works because forming such intentions triggers the construction of individual plans to realize part of a shared activity; and this leads group members to behave in ways that bring about the shared end (cf., Bratman, 2014).

Something similar happens when groups make temporally extended plans. Consider a group that has a small pot of money to invest. They settle on an initial plan, and occasionally meet to update that plan in light of their successes and failures. This group should be able to revise or abandon their plan if things go badly, so long as relevant information is shared and available. But group members routinely overcommit to ends they have settled upon

---

10. There are many ways for groups to entrench power and hierarchy and to cause collaboration to collapse in favor of patterns of exploitation and deference. In many groups, polarization effects will arise, and conformity, hierarchy, and prestige biases will often prevent the emergence of egalitarian group structures. I focus here on the best-case scenario, much as Bratman (2014) does. I think that work carried out by Elinor Ostrom (1990) demonstrates that these sorts of cases are possible; but they are difficult to sustain, and the emergence of any pattern of exploitation can prevent a group from capitalizing on its informational advantages.

as a group, and they often overestimate their chances of success (Wieber et al., 2012, p. 278). As with individuals who make similar errors, this is the result of moving from a deliberative to an implementation mindset. And here too, performance can be improved with implementation intentions. When the success of a project declines and requires lowering investments, groups that intend to judge a project as onlookers who aren't responsible for earlier decisions often fail to adjust their plan. As observers, they should be able to disengage; but they don't. By contrast, groups whose members form an implementation intention ("When we are about to make an investment decision, then we will judge the project as onlookers who are not responsible for earlier decisions!") adjust their investments to track their current situation. In both cases, people are reflexively led to act as participants in a joint activity; but those who use implementation intentions can exploit cognitive strategies that allow them to take advantage of the information that is distributed among group members (Wieber, Thürmer, & Gollwitzer, 2013). By recognizing the potential for epistemic errors, they can use their foreknowledge as group members to guide their joint activity.

Nonetheless, the same limitations remain. As Dennett and Bratman both argue, planning to do something together allows us to precommit to actions that are consistent with our shared goals and values. And where people think and act together, they can capitalize on shared information to successfully execute joint actions, so long as they are mutually responsive. But this only works where we already have strong commitments to pursuing a particular goal; like individual intentions, joint intentions can only link plans to currently salient information, and they can only facilitate top-down control over our behavior. This opens up more degrees of freedom, but it doesn't explain how we can move beyond the thoughts that readily occur to us.

## 8. Shared Plans

I contend that deliberation and planning can be realized by distributed networks of individuals, who use discrete representations to broadcast their own attitudes and revise them, in parallel, in light of shared modes of thinking. This can happen where cooperating individuals convey just enough information to coordinate, using "trading languages" that allow them to query each other while ignoring many irrelevant details (Galison, 1997, p. 883). As a result, acts of planning together become transactive; we rely on information that has been provided by others, and take part in the collaborative construction of prospective representations that can guide our joint activities. At the same time, this process is guided by subpersonal mechanisms that facilitate alignment in dynamic interpersonal

contexts.<sup>11</sup> This is a complex and contentious idea, so let's move through it with a familiar example before turning to the power of this process to create new degrees of freedom.

Imagine a long-term couple who take a vacation together. Over the years, they have learned that one of them will know where the best coffee is, while the other will know where the best museums are (with substantial redundancy in other domains, and some redundancy even here). When they arrive in a city, one of them might focus on getting to their favorite café, while the other might think about heading to a modern art exhibit. Both of them may have an initial inclination to nudge the other, as both will assume that their own plan is more feasible and desirable. But suppose they have also developed strategies for navigating disagreement, and so they decide to talk about what to do first. They might realize that getting lunch first will allow them to make further decisions without arguing—something neither of them would have noticed before. And by verbally interacting, and attending to patterns of nonverbal behavior, they can construct a shared plan that will account for their independent needs and shared goal of enjoying their vacation. In this respect, their capacities parallel the capacities of couples that can reconstruct shared memories together (Huebner 2016; Theiner, 2013; Tollefsen, Dale, & Paxton, 2013), and thinking briefly about memory can shed light on the nature of this constructive process.

Like plans, episodic memories are constructed by fleshing out the details of skeletally structured representations (Bartlett, 1932; De Brigard, 2014; Neisser, 1981). We “draw on the elements and gist of the past, and extract, recombine and reassemble them into imaginary events that never occurred in that exact form” (Schacter & Addis, 2007, p. 27). And like plans, the construction of memories often draws on information that is anchored in the material and social structure of our world (Hutchins, 2005). Long-term couples can often minimize the demands on limited cognitive resources by storing different kinds of memories (often with substantial redundancy); this allows them to increase the breadth of their knowledge, while simultaneously increasing its depth, using a virtual

---

11. Thanks are due to John Sutton for pushing me on this point in another context. The details of my argument here are articulated more fully in Huebner (2014). But the point runs much deeper than I have time to address in this chapter (which is already much too long). As Maureen Sie (personal communication, October 8, 2015) notes, I haven't said much of anything about the role of our moral reactive attitudes in guiding collaborative and collective behavior. She is right to flag this, and any plausible attempt to fill out the ideas that I have sketched here will have to acknowledge the importance of these attitudes to sustaining, undermining, and opening up new possibilities for collaborative action. On this point, I agree with much of the argument that is developed in Sie (2013), and in my future attempts to expand on this perspective I plan to return to these issues in more detail.

memory store that spans the transactive network that they constitute (Wegner, 1995). Such couples then construct representations of past events by dynamically adjusting and recalibrating their individual mental states to track the explicit representations that they broadcast to one another. As a result, they can often reconstruct more detailed and elaborate memories by cross-cuing one another and engaging in conversation; and their conversations allow them to recall details of past events that neither would have remembered on their own (Harris, Keil, Sutton, Barnier, & McIlwain, 2011). The process of remembering thus relies on shared computations over linguistic and gestural representations, and the constructive process only needs to be carried out once—in conversation. Transactive memories emerge because couples “think out loud,” and externalize the processes of remembering; by cross-cuing one another, they implement an interfaced system, distributed memories bind the group together, and “any one individual is incomplete without being able to draw on the collective knowledge of the rest of the group” (Wegner & Ward, 2013, p. 58). In essence, they form a flexibly coupled system that mimics the architecture of the BitTorrent protocol.

Likewise, when couples plan together, they can retrieve and broadcast individually stored representations, using a process of cross-cuing to construct a shared plan that doesn’t need to be represented prior to conversation. By constructing a plan together, they implement an interfaced system that produces shared representations, which guide collaborative behavior, and generate individual expectations that are shared by individuals qua group members, just as Bratman suggests. As they flesh out the details of their plan, facts that are crucial to their current interactive context may arise in ways that neither of them would have considered on their own; consequently, they might act in ways that diverge from the preferred option that either would have selected, but that are well suited to their interactive context. Domain-general computations over linguistic and gestural representations are used to construct shared plans, and there is no reason for this process to occur more than once. By “thinking out loud,” people can thereby externalize the deliberative process, yielding explicit representations that are relevant to the guidance of collaborative behavior, even though the implementation of the resulting plans remains an individual process.

We can plan together as group members who share a great deal of epistemic common ground, and we often do so. But this can create epistemic echo chambers, where ideas are replicated, sustained, and more deeply entrenched.<sup>12</sup> The regulative dimensions of mindshaping have precisely this effect. They make it easier for us to accept habitual patterns of thought and action, and easier for us to

---

12. I borrow this idea of an epistemic echo chamber from Benjamin Elzinga.



passively accept the power structures that arise through processes of cultural evolution (Dennett 2017). We are inclined to act in accordance with social norms, and as a result we often find ourselves on local peaks in an adaptive landscape, happy enough with our situation, and unable to imagine other ways that the world can be. This isn't always a bad thing. Those of us who benefit from local power structures tend to act in ways that feel right to us. And this shouldn't be a surprise; Darwinian evolution tends to be a stabilizing force, which preserves traits that help animals occupy the available niches. But we can construct novel niches, because we can imagine possibilities that are better than the ones that we have come to expect.

## 9. Help From Our Friends

The forms of joint deliberation and joint agency I have just been discussing each play a critical role in the process of imagining another world, and both must be in place to open up and sustain novel forms of elbow room. Joint deliberation can help us realize that our current ways of thinking have emerged as a result of our contingent learning histories, and it can help us find new ways to think about the world we inhabit. This happens as we express facts about our notional world with explicit representations, and submit them to practices of evaluation and revision; and where shared deliberations arise outside of epistemic echo chambers, when we interact with people who have different expectations and different insights about aspects of our shared practices, this can be a highly productive process. Specifically, when we *listen* to people who have different forms of embodiment than us, or different backgrounds that lead them to expect different things, we can come to see the world in very different ways.

In a strange way, I think that Dennett hit on a similar insight in one of his early attempts to address questions about freedom and responsibility. I paraphrase here, replacing Dennett's (1978, p. 295) claims about subpersonal processes with claims about suprapersonal processes:

My model of decision-making has the following feature: when we face a normatively significant decision, we can often rely on others to generate considerations that we wouldn't have access to on our own. Which considerations arise will not be fully determined by the fact that we are planning together; and as we evaluate these considerations, some will be immediately rejected as irrelevant, or as inconsistent with our shared goals. The considerations we select as having a reasonable bearing on our shared behavior will then be submitted to practices of giving and asking for reasons, and if we are mutually responsive, and non-exploitative, the



considerations we arrive at will ultimately serve as predictors and explicators of our decisions and actions *qua* group members.

It should come as no surprise that this is a common thread that runs through the collaborative practices of early-20th-century anarchists, the social justice work of Óscar Romero and Paul Farmer, and the attempts by the Zapatistas of Chiapas to build a better world (Lynd, 2012; Lynd & Grubačić, 2008). Each group puts *listening* at the core of their radical project. They see that it is often only by embedding ourselves in a world that we don't quite understand that we can begin to change the options that we see, and begin to see the unfounded presuppositions that we have signed onto unreflectively. By attempting to understand the world as others do, we can sometimes recognize the contingency of things that have seemed necessary.<sup>13</sup>

Revealing contingencies, however, is only the first step in breaking down the teleological assumptions that lead people to believe that our social arrangements must continue to be what they currently are; and it is the first step in making more elbow room for new forms of action through acts of planning together. To the extent that we embed ourselves in shared practices that are designed to foster mutual understanding, we may be able to uncover options that we would have missed on our own; and some of these ideas may push us beyond those that would have been available given our reinforcement history. I think Dennett knows this, but I have no idea whether he knows that he knows it (or even whether he believes it).

Specifically, I contend that a form of mental contrasting can be useful for increasing our freedom. By imagining a desired future, and reflecting on facts that stand in the way of reaching it, individuals can enhance the cognitive relevance of the desirable features of an action, and highlight the feasibility of particular plans (Oettingen & Gollwitzer, 2010). But I contend that the reliance of this practice on explicit representations should also make a collaborative form of mental contrasting possible. If so, this should be able to open up an inherently social form of elbow room. Achieving this, however, is not easy. To see this, we only need to reflect on the shared habits that emerge in long-term relationships. Breaking out of these habits often requires interjecting new forms of thinking, and developing new habits of communication, to allow previously undisclosed facts to be brought to the fore. Teaching individuals to do this is big business, and not a particularly successful business at that. It requires cultivating new forms of thinking,

---

13. This is one reason why it makes sense to teach all of the world's religions in schools (Dennett, 2006).

as well as abilities to think across different social and cultural frameworks, and the willingness to interact with people from other professional groups, socio-economic statuses, religions, political persuasions, and more (Gavazzi, personal communication, September 15, 2015). Strategies for doing this will only become more difficult to enact as groups become larger and more diverse.

Nonetheless, there is reason to believe that if individuals become comfortable articulating their conceptions of a desired future, and collaboratively reflect on the facts about the world that stand in their way, they will be able to enhance the cognitive relevance of the desirable features of novel patterns of behavior, and highlight the feasibility of the plans they have constructed as a group.<sup>14</sup> Having articulated a shared idea about which ends to pursue, groups of people can begin to set out plans that will allow them to solidify new ways of living and acting together. Indeed, people often use forms of social and self-monitoring to sustain forms of collective action, and they are most likely to succeed where strategies for managing defection and cooperation are self-organized, and grounded in ideals that everyone adopts (Ostrom, 1990). Where people agree that an issue they face is important, retain a distributed and collective form of autonomy over mutually agreed upon rules, and develop community-centered practices for monitoring and sanctioning others, they can act to foster ongoing forms of collaboration. Again, this isn't easy, but people can work toward this end by precommitting to particular practices *qua group members*, and thereby decreasing the likelihood of defection.

Of course, we must first find ways to sustain forms of thinking that allow us to act as equals, especially where people dominate discussions or engage in exploitative practices (Bratman, 2014). In a world structured by exclusionary practices, it will be difficult to sustain non-exploitative practices and non-dominating forms of speech. Future-directed intentions can be valuable in this context, as can precommitments that impose normative pressure against those who defect from shared practices. As group members, pressure isn't just psychological; it's also social. And by changing our patterns of social engagement, we can begin to pull ourselves toward counternormative, yet preferable, ideals with help from our friends.

Finally, planning together can create novel environmental contingencies, which can prevent the forms of backsliding to which Bayesian agents are susceptible. The

---

14. One intriguing piece of data in this regard is reported by Woolley, Chabris, Pentland, Hashmi, and Malone (2010), who found that the performance on a wide range of tasks, for people working in small groups of two to five people, is not predicted by the average or maximum individual intelligence of group members, but is highly correlated with things like the average social sensitivity of group members, equitable turn-taking, and the proportion of women in a group.

structure of the cognitive prosthetics we rely upon has an enormous impact on the ways that we update ongoing behavior. We attune to social practices, and our high-level attitudes and low-level reactions tend to entrain to the local patterns we encounter, leading to stable institutional structures against which future attitudes and behavior attune. By building stable microworlds that accord with our forward-looking expectations, we can get the feedback loop that I discussed earlier to solidify ways of thinking and acting that are consistent with our values and ideals, instead of allowing the world to undermine them. Collaborative actions that are grounded in prefigurative imagination will be met with evaluative feedback suggesting that we are acting rightly. As a result, new hypotheses can be sustained, which contradict dominant forms of social power; and as the brain searches for the linked set of hypotheses that make incoming data most plausible, our expectations will then shift toward the local patterns of interaction we are in the process of constructing. To my mind, this seems like the kind of freedom that is most worth wanting: it is the political freedom to change social norms, by collaboratively resisting them in the short run, and entrenching them in shared social practices that open up novel degrees of freedom.<sup>15</sup>

## 10. Epilogue

There is a great deal more work to be done in fleshing out a plausible, socially situated view of agency. And my claims in this chapter are probably too optimistic, given the pervasive role of exclusionary ideals and hierarchical ideologies in the world we inhabit. So for now, I just want to note that we should not be content with the kinds of freedom we can pursue on our own. Since the human brain is a predictive machine, we are constrained by social and environmental contingencies. But the point of doing philosophy isn't simply to understand *that* we are constrained, it is to find ways of changing *how* we are constrained. We are rabid

---

15. People who live in more diverse communities, and who interact with the members of other racial groups in a more diverse range of situations, for example, tend to be less racially biased, and tend to have explicit attitudes that are more egalitarian (Dasgupta & Rivera, 2008). Inhabiting such neighborhoods creates and reinforces positive implicit associations, which can counteract the biases that structure the rest of our world (Dasgupta, 2013, p. 247). Living in such neighborhoods also helps mitigate the appeal of colorblind ideologies, and heightens the awareness of forms of structural racism that go beyond explicitly racist attitudes (Dasgupta, 2013; Garner, 2007, pp. 45–46). Such attitudes can help us to see the hegemonic ideology of Whiteness as contingent, distorting, and dispensable, instead of seeing it as a necessary conceptual framework. Of course, power is never given up easily, and there are many opportunities to abandon antiracist attitudes in favor of the comfort of White ideology. But diverse spaces, structured around diverse goals and values, may help to provide a place for anti-racist ideologies and critical approaches to Whiteness to develop. For a slightly more robust discussion of these issues, see Huebner (2016).

social-niche constructors, and another world is possible. But acting in ways that go beyond our contingent learning histories requires planning together, imagining together, and acting together in ways that prefigure the world we would like to inhabit. In the process, our own preferences will probably need to change, and they are only likely to do so as a result of accompanying others in new kinds of shared practices.

When we act on our own, we will be lucky if our behavior happens to align with values and ideals that we can reflectively avow. But this is not guaranteed, even where things feel relatively comfortable. By contrast, where we collaboratively build social structures, our values and judgments can stabilize around positive and productive biases, which will become calcified in social norms and practices, and guide ongoing behavior as free-floating rationales. Put somewhat differently, planning and acting together can yield a form of social-niche construction, which is grounded in our capacity to think and act together; as associative thinkers, the ends we seek become more stable as we surround ourselves with others who value the same things as we do (Spinoza, 1677/2002, p. 339). The Cartesian approach to “free will” that tends to dominate discussions of agency focuses attention on ways of changing ourselves and of making ourselves better people. As naturalistic philosophers, we should recognize that we are often nudged around by interactions with the world in ways that we cannot control, but as groups working together, we can collectively devise circumstances that enhance our collective power to act. So the ethical question should always be: How can we construct forms of collective action that open up possibilities that we don’t possess as individuals. And this makes ethics, as well as reflections on agency and freedom, a matter of politics, not metaphysics.

## Acknowledgments

The ideas in this paper developed over the course of several helpful conversations with Gunnar Björnsson, Justin Caouette, Gregg Caruso, Mattia Gallotti, Joey Jebari, Pete Mandik, Manuel Vargas, Tad Zawidzki, and of course, Dan Dennett. John Gavazzi, Ruth Kramer, Rebecca Kukla, and Maureen Sie each read a complete draft and offered insightful comments on where I had gone wrong. I think it’s fair to say that I got by with quite a bit of help from my friends

## Works Cited

- Adriaanse, M. A., Gollwitzer, P. M., De Ridder, D. T., de Wit, J. B., Kroese, F. M. (2011). Breaking habits with implementation intentions. *Personality and Social Psychology Bulletin*, 37, 502–513.

- Adriaanse, M. A., Vinkers, D. D., De Ridder, D. T. Hox, J. J., de Wit, J. B. (2011). Do implementation intentions help to eat a healthy diet? *Appetite*, 56, 183–193.
- Ainslie, G. (2001). *Breakdown of will*. Cambridge, England: Cambridge University Press.
- Akins, K. (1996). Of sensory systems and the “aboutness” of mental states. *Journal of Philosophy*, 93, 337–372.
- Barrett, L. F. (2014). The conceptual act theory: A précis. *Emotion Review*, 6, 292–297.
- Barrett, L. F. (2015, April 17). When a gun is not a gun. *New York Times*. Retrieved from <http://goo.gl/PvXLDt>
- Bartlett, F. (1932). *Remembering*. Cambridge, England: Cambridge University Press.
- Belanger-Gravel, A., Godin, G., & Amireault, S. (2013). A meta-analytic review of the effect of implementation intentions on physical activity. *Health Psychology Review*, 7, 23–54.
- Blomberg, O. (2015). “Review of Shared Agency,” *Analysis*, 75(2): 346–348.
- Botvinick, M., & Toussaint, M. (2012). Planning as inference. *Trends in Cognitive Sciences*, 10, 485–588.
- Brandstätter, V., & Frank, E. (2002). Effects of deliberative and implemental mindsets on persistence in goal-directed behavior. *Personality and Social Psychology Bulletin*, 28, 1366–1378.
- Bratman, M. (1987). *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.
- Bratman, M. (2008). Intention, belief, practical, theoretical. In S. Robertson (Ed.), *Spheres of reason* (pp. 29–51). Oxford, UK: Oxford University Press.
- Bratman, M. (2014). *Shared agency: A planning theory of acting together*. New York, NY: Oxford University Press
- Christman, J. (2004). Relational autonomy, liberal individualism, and the social constitution of selves. *Philosophical Studies*, 117, 143–164.
- Clark, A. (1996). Linguistic anchors in the sea of thought? *Pragmatics and Cognition*, 4(1), 93–103.
- Coleman, N. A. T. (2015, August 24). How philosophy was “whitewashed.” [Web blog post]. Retrieved from <http://goo.gl/cCypbS>
- Couzin, I. D. (2009). Collective cognition in animal groups. *Trends in Cognitive Sciences*, 13, 36–43.
- Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, 17, 363–366.
- Crockett, M. J., Braams, B. R., Clark, L., Tobler, P. N., Robbins, T. W., & Kalenscher, T. (2013). Restricting temptations. *Neuron*, 79, 391–401.
- Dasgupta, N. (2013). Implicit attitudes and beliefs adapt to situations: A decade of research on the malleability of implicit prejudice, stereotypes, and the selfconcept. In P. G. Devine & E. A. Plant (Eds.), *Advances in experimental social psychology* (Vol. 47, pp. 233–279). London: Academic Press.
- Dasgupta, N., & Rivera, L. M. (2008). When social context matters: The influence of long-term contact and short-term exposure to admired outgroup members on implicit attitudes and behavioral intentions. *Social Cognition*, 26, 54–66.

- De Brigard, F. (2014). Is memory for remembering? Recollection as a form of episodic hypothetical thinking. *Synthese*, 191, 155–185.
- Dehaene, S., & Changeux, J. P. (2000). Reward-dependent learning in neuronal networks for planning and decision making. *Progress in Brain Research*, 126, 217–229.
- Dennett, D. C. (1978). On giving libertarians what they say they want. In D. C. Dennett, *Brainstorms: Philosophical essays on mind and psychology*. Cambridge, MA: MIT Press.
- Dennett, D. C. (1995). *Darwin's dangerous idea*. New York, NY: Simon and Schuster.
- Dennett, D. C. (2003). *Freedom evolves*. London: Penguin.
- Dennett, D. C. (2006). *Breaking the spell: Religion as a natural phenomenon*. London: Penguin.
- Dennett, D. C. (2015). Why and how does consciousness seem the way it seems? Open MIND: 10(T). Frankfurt am Main: MIND Group. doi:10.15502/9783958570245
- Dennett D. C. (2017). *From bacteria to Bach and back*. New York, NY: W. W. Norton.
- Drayson, Z., & Clark A. (forthcoming). Augmentation, agency, and the spreading of the mental state.
- Fitch, W., & Martins, M. D. (2014). Hierarchical processing in music, language, and action: Lashley revisited. *Annals of the New York Academy of Sciences*, 1316, 87–104.
- Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, 13, 293–301.
- Gallotti, M., & Frith, C. (2013). Social cognition in the we-mode. *Trends in Cognitive Sciences*, 17, 160–165.
- Galison, P. (1997). *Image and logic*. Chicago, IL: University of Chicago Press
- Garner, S. (2007). *Whiteness: An Introduction*. London: Routledge.
- Gilbert, S. J., Gollwitzer, P. M., Cohen, A. L., Burgess P. W., & Oettingen, G. (2009). Separable brain systems supporting cued versus self-initiated realization of delayed intentions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 905–915.
- Gollwitzer, P. (1999). Implementation intentions. *American Psychologist*, 54, 493–503.
- Gollwitzer, P. (2012). Mindset theory of action phases. In P. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (Vol.1, pp. 526–545). London: Sage.
- Gollwitzer, P. M. (2014). Weakness of the will: Is a quick fix possible? *Motivation and Emotion*, 38, 305–322.
- Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in Experimental Social Psychology*, 38, 69–119.
- Gross, Z. (2012, June 4). Navigation and disabled interdependence. [Web blog post]. Retrieved from <http://goo.gl/s8iQai>
- Harris, C., Keil, P., Sutton, J., Barnier, A., & McIlwain, J. (2011). We remember, we forget. *Discourse Processes*, 48(4), 267–303.

- Hohwy, J. (2013). *The predictive mind*. Oxford, UK: Oxford University Press.
- Holroyd, J. (2011). The metaphysics of relational autonomy. In C. Witt (Ed.), *Feminist metaphysics* (pp. 99–115). Dordrecht, The Netherlands: Springer.
- Holton, R. (2009). *Willing, wanting, waiting*. New York, NY: Oxford University Press.
- Huebner, B. (2014). *Macrocognition*. New York, NY: Oxford University Press.
- Huebner, B. (2016). Implicit bias, reinforcement learning, and scaffolded moral cognition. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy: Metaphysics and Epistemology* (Vol. 1, pp. 47–79). Oxford, UK: Oxford University Press.
- Huebner, B. (2016). Transactive memory reconstructed. *The Southern Journal of Philosophy*, 54(1), 48–69.
- Humphreys, G., Forde, E., & Riddoch, M. (2001). The neuropsychology of everyday actions. In B. Rapp (Ed.), *The handbook of cognitive neuropsychology* (pp. 565–592). Cambridge, MA: MIT Press.
- Hutchins, E. (2005). Material anchors for conceptual blends. *Journal of Pragmatics*, 37, 1555–1577.
- Jackendoff, R. (2007). *Language, consciousness, culture: Essays on mental structure*. Cambridge, MA: MIT Press.
- Kim, M., Loucks, R., Palmer, A., Brown, A., Solomon, K., Marchante, A., & Whalen, P. (2011). The structural and functional connectivity of the amygdala: From normal emotion to pathological anxiety. *Behavioural Brain Research*, 223, 403–410.
- Kirsh, D. (2006). Distributed cognition: A methodological note. *Pragmatics and Cognition*, 14, 249–262.
- Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, 18, 513–549.
- Klucharev, V., Hytonen, K., Rijpkema, M., Smidts, A., & Fernandez, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron*, 61, 140–151.
- Klucharev, V., Munneke, M. A., Smidts, A., & Fernandez, G. (2011). Downregulation of the posterior medial frontal cortex prevents social conformity. *Journal of Neuroscience*, 31, 11934–11940.
- Kosslyn, S. (2006). On the evolution of human motivation. In S. Platek, T. Shackelford, & J. Keenan (Eds.), *Evolutionary cognitive neuroscience* (pp. 541–553). Cambridge, MA: MIT Press.
- Kukla, R. (2005). Conscientious autonomy: Displacing decisions in health care. *Hastings Center Report*, 35(2), 34–44.
- Kutz, C. (2000). Acting together. *Philosophy and Phenomenological Research*, 61(1): 1–31.
- Landsiedel, J., & Gilbert, S. J. (2015). Creating external reminders for delayed intentions: Dissociable influence on “task-positive” and “task-negative” brain networks. *NeuroImage*, 104, 231–240.
- Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112–146). New York, NY: Wiley.



- Lynd, S. (2012). *Accompanying: Pathways to social change*. Oakland, CA: PM Press.
- Lynd, S., & Grubačić, A. (2008). *Wobblies and Zapatistas: Conversations on anarchism, marxism and radical history*. Oakland, CA: PM Press.
- MacKenzie, C. (2014). Autonomy. In J. Arras, E. Fenton, & R. Kukla (Eds.), *The Routledge companion to bioethics* (pp. 277–290). New York, NY: Routledge.
- Mackenzie, C., & Stoljar, N. (Eds.). (2000). *Relational autonomy: Feminist perspectives on autonomy, agency and the social self*. New York, NY: Oxford University Press, 277–290.
- Mahon, B., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology-Paris*, 102(1), 59–70.
- McLeod, C. (2002). *Self-trust and reproductive autonomy*. Cambridge, MA: MIT Press.
- Mendoza, S. A., Gollwitzer, P. M., & Amodio, D. M. (2010). Reducing the expression of implicit stereotypes. *Personality and Social Psychology Bulletin*, 36, 512–523.
- Montague, R. (2006). *Why choose this book?* New York, NY: Dutton.
- Neisser, U. (1981). "John Dean's memory." *Cognition*, 9, 1–22.
- Oettingen, G., & Gollwitzer, P. (2010). Strategies of setting and implementing goals: Mental contrasting and implementation intentions. In J. E. Maddux & J. P. Tangney (Eds.), *Social psychological foundations of clinical psychology* (pp. 114–135). New York, NY: Guilford.
- Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge, England: Cambridge University Press.
- Pacherie, E. (2006). Toward a dynamic theory of intentions. In S. Pockett, W. P. Banks, & S. Gallagher (Eds.), *Does consciousness cause behavior?* (pp. 145–167). Cambridge, MA: MIT Press.
- Polanía, R., Moisa, M., Opitz, A., Grueschow, M., & Ruff, C. C. (2015). The precision of value-based choices depends causally on fronto-parietal phase coupling. *Nature Communications*, 6. Online publication. Retrieved from doi.10.1038/ncomms9090
- Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, 43, 151.
- Rosenbaum, D. A., Cohen, R. G., Jax, S. A., Weiss, D. J., & Van Der Wel, R. (2007). The problem of serial order in behavior: Lashley's legacy. *Human Movement Science*, 26, 525–554.
- Schelling, T. C. (1966). *Arms and influence*. New Haven, CT: Yale University Press.
- Schacter, D., & Addis, D. (2007). The cognitive neuroscience of constructive memory. *Philosophical Transactions of the Royal Society B*, 362, 773–786.
- Schwabe, L., & Wolf, O. T. (2013). Stress and multiple memory systems. *Trends in Cognitive Sciences*, 17, 60–68.
- Seligman, M., Railton, P., Baumeister, R., & Sripada, C. (2013). Navigating into the future or driven by the past. *Perspectives on Psychological Science*, 8(2), 119–141.



- Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, 18, 186–193.
- Sie, M. (2013). Free will an illusion? An answer from a pragmatic sentimentalist point of view. In G. Caruso (Ed.), *Exploring the illusion of free will and moral responsibility* (pp. 273–289). Lanham, MD: Lexington Books.
- Sie, M. (2014). Self-knowledge and the minimal conditions of responsibility. *Journal of Value Inquiry*, 48(2), 271–291.
- Simon, H. A. (1969). *The sciences of the artificial*. Cambridge, MA: MIT Press.
- Spinoza, B. (1677/2002). *The complete works* (S. Shirley & M. Morgan, Eds.). New York, NY: Hackett.
- Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333, 776–778.
- Táiwò, O. (in prep). *Why we are not what we seem: The social ascription critique of agency*.
- Taylor, S. E., & Gollwitzer, P. M. (1995). Effects of mindset on positive illusions. *Journal of Personality and Social Psychology*, 69, 213–226.
- Theiner, G. (2013). Transactive memory systems: A mechanistic analysis of emergent group memory. *Review of Philosophy and Psychology*, 4, 65–89.
- Thürmer, J., Wieber, F., & Gollwitzer, P. (2014). When unshared information is the key. *Journal of Behavioral Decision Making*, 28, 101–113. doi:10.1002/bdm.1832
- Tollefsen, D. (2005). Let's pretend: Children and joint action. *Philosophy of the Social Sciences*, 35(1): 75–97.
- Tollefsen, D. (2014). A dynamic theory of shared intention. In S. Chant & F. Hindriks (Eds.), *From individual to collective intentionality: New Essays* (pp. 13–33). Oxford, UK: Oxford University Press.
- Tollefsen, D., & Dale, R. (2012). Naturalizing joint action. *Philosophical Psychology*, 25, 385–407.
- Tollefsen, D., Dale, R., & Paxton, A. (2013). Alignment, transactive memory, and collective cognitive systems. *Review of Philosophy and Psychology*, 4, 49–64.
- Webb, T., Sheeran, P., & Pepper, A. (2012). Gaining control over responses to implicit attitude tests. *British Journal of Social Psychology*, 51(1): 13–32.
- Wegner, D. (1995). A computer network model of human transactive memory. *Social Cognition*, 13, 1–21.
- Wegner, D., & Ward, A. (2013, December 1). The internet has become the external hard drive for our memories. *Scientific American*, 309(6). Retrieved from <http://goo.gl/vEB1h2>
- Wieber, F., Thürmer, J., & Gollwitzer, P. (2012). Collective action control by goals and plans. *American Journal of Psychology*, 125, 275–290.
- Wieber, F., Thürmer, J., & Gollwitzer, P. (2013). Intentional action control in individuals and groups. In G. Seebaß, M. Schmitz, & P. M. Gollwitzer (Eds.), *Acting intentionally and its limits* (pp. 133–162). Berlin, Germany: De Gruyter.

- Wilson-Mendenhall, C. D., Barrett, L. F., Simmons, W. K., & Barsalou, L. W. (2011). Grounding emotion in situated conceptualization. [Supplemental Material]. *Neuropsychologia*, 49, 1105–1127.
- Wilutzky, W. (2015). Emotions as pragmatic and epistemic actions. *Frontiers in psychology*, 6, 1593.
- Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, 330, 686–688.

## 10.2 REFLECTIONS ON BRYCE HUEBNER

Daniel C. Dennett

Bryce Huebner agrees with me that we humans, unlike other animals, have the ability, thanks to language, to form and criticize explicit plans, and that this talent equips us with powers of expectation and self-control far exceeding those of other species. This major innovation in cognitive architecture is, moreover, what chiefly distinguishes us as candidates for moral agency. I have made this a central theme in my work since “Conditions of Personhood” (1976); we may not be cognitive angels, but our imperfect capacity to be moved by represented reasons expands our moral imaginations, reliably furnishing our decision-making efforts with the appropriate considerations so that we can fairly be held accountable for doing the right thing. Huebner gets it, and expands it nicely by introducing Gollwitzer’s distinction between the *deliberative* and *implementation* mindsets. This work is new to me, and by adding crucial details it cranks up the power of the intentional stance to account for biases built into our rationality. The thought processes of any agent can be idealized as asking, over and over, “what should I do now?” (an idealization that has many clear instantiations in computer programs that are richly equipped with self-monitoring loops), and what Gollwitzer draws attention to is that these incessant questions fall naturally into two versions: the “all things considered” questions about which large projects to engage in (and whether to abandon an ongoing project in favor of something else), and the more narrowly focused questions about which means of accomplishing an assumed goal are apt to be successful. This makes sense, Huebner notes, “but this leads people to be more optimistic about their chances of success, and more partial in their judgments about the desirability of the ends they pursue” (p. 298). Putting on temporary blinders speeds our implementation decisions, but we pay a price for the oversimplification.

Another aspect of our implementation mindset is that it can be initiated by making a request of ourselves, an act that might look foolishly redundant were it not the fact that once we have undertaken to fulfill the self-request, this elevates the moral importance of the project. “Like failures to comply with accepted requests, the failure to follow through on our commitments feels bad, and calls for a justification” (p. 298). (Note that all this requesting, promising oneself, setting up cues, and self-monitoring must be conscious to be effective. See the discussion of word-stem completion in my “Reflections on David Rosenthal,” chapter 5.2, this volume.) By explicitly tying a policy to our self-image as a good and rational person, we can introduce benign manipulations into our cognitive architecture, sometimes with striking effects. Huebner relays the surprising and heartening discovery by Mendoza and Webb among others that such “implementation intentions can moderate the perceptual tendency to mistake a cellphone for a weapon in a racialized context” (p. 300).

“But as powerful as they are, implementation intentions only guild behavior where we already have strong commitments to particular goals” (p. 300). This observation leads Huebner to be unpersuaded by my rosy conclusion that we are the “authors and executors of these policies, even though they are compiled from parts we can only indirectly monitor and control” (Dennett 2003, p. 239). Where I say we are “capable of altering course at any point, abandoning goals, switching allegiances, forming cabals and then betraying them, and so forth” (Dennett 2003, p. 155), Huebner notes:

But we can only do so on the basis of goals and values that we have acquired through our interactions with the world. And where problematic patterns are pervasive, and where we act in ways that feel right because they are statistically regular, we will only hit upon normatively valuable practices by accident or luck. (p. 306)

Like boats whose courses are constantly warped by the tidal currents in which they are moving, our Bayesian brains have no way of filtering out the biases that are ambient or dominant in a culture; indeed, our brain *tracks* the biases with high fidelity, just the way the boat’s “course made good over the ground” tracks the currents in which it moves. We are therefore somewhat at the mercy of the culture we are immersed in; truly radical breaks with tradition are hard to countenance, even when our reason-giving competence vividly supports them. Is there any way of overcoming this cultural myopia?

Huebner sees that there is some hope, if we adopt the practice of getting a little help not just from our friends but from those we find most different and

unfathomable. A process of *group* self-monitoring can be substantially improved, it turns out, by adopting an explicit policy: “When we are about to make the final decision, then we will go over the advantages of the non-preferred alternatives again” (p. 313). Huebner then ingeniously takes an old passage of mine about subpersonal processes and turns it into a description of the superpersonal collaborative practices of “early 20th century anarchists, the social justice work of Óscar Romero and Paul Farmer, and the attempts by the Zapatistas of Chiapas to build a better world” (p. 318). An unimagined application indeed! But it does fit beautifully with the vision he is articulating and defending here:

Revealing contingencies, however, is only the first step in breaking down the teleological assumptions that lead people to believe that our social arrangements must continue to be what they currently are; and it is the first step in making more elbow room for new forms of action through acts of planning together. To the extent that we embed ourselves in shared practices that are designed to foster mutual understanding, we may be able to uncover options that we would have missed on our own; and some of these ideas may push us beyond those that would have been available given our reinforcement history. I think Dennett knows this, but I have no idea whether he knows that he knows it (or even whether he believes it). (p. 318)

I know it now. And I agree wholeheartedly with Huebner’s conclusion that

the kind of freedom that is most worth wanting . . . is the political freedom to change social norms, by collaboratively resisting them in the short run, and entrenching them in shared social practices that open up novel degrees of freedom . . . we should not be content with the kinds of freedom we can pursue on our own . . . And this will make ethics, as well as reflections on agency and freedom, a matter of politics, not metaphysics. (pp. 320–321)

## Works Cited

- Dennett, D. C. (2003). *Freedom evolves*. London: Penguin.
- Dennett, D. (1976). Conditions of personhood. In A. O. Rorty (Ed.), *The identities of persons* (pp. 175–196). Berkeley: University of California Press.

# 11.1

## DENNETT ON *BREAKING THE SPELL*

Lynne Rudder Baker

Daniel C. Dennett is an important professional philosopher, one of the most influential of his generation. He is also a public intellectual who writes broadly about science and religion. Dennett is often identified as one of the “Four Horsemen of the New Atheism”—the other three being Richard Dawkins, Sam Harris, and Christopher Hitchens—but Dennett is much more careful and serious than the other three. (For example, Dennett’s discussions of Jared Diamond, Pascal Boyer, Scott Atran, Dan Sperber, and David Sloan Wilson, *inter alia*.)

Although Dennett, like the other Horsemen, is a confident atheist, his most sustained work on religion, *Breaking the Spell*, has the sober subtitle, “Religion as a Natural Phenomenon.” *Breaking the Spell* is a lively and engaging book. It is full of tidbits of general interest, of reflections on liberal democracy, of weird examples from biology. (See, for example, the case of the sea squirt, an organism that uses a rudimentary nervous system for locomotion in seeking a rock to cling to for life and then disassembles and assimilates its own nervous system when it is no longer needed; Dennett 2006, p. 366.) Indeed, *Breaking the Spell* is brimming with fascinating and controversial ideas, some more successful (in my opinion) than others.

Let’s begin by asking, in *Breaking the Spell*, what spell does Dennett want to break? The first spell, the one that he thinks urgently needs to be broken, is the one that surrounds religion with a taboo “against a forthright, scientific, no-holds-barred investigation of religion as one natural phenomenon among many” (Dennett, 2006, p. 17). Breaking that spell risks also breaking another and more important spell: “the life-enriching enchantment of religion itself” (p. 17). Dennett’s main aim is to break the spell that removes religion from the reach of science; with regard to the second and more important spell—that of the enchantment of religion itself—Dennett sometimes seems willing

to let the chips fall where they may. (But at other times, he gives the chips a little nudge; see Dennett 2006, p. 174.)

Dennett gives a provisional definition of religions: Religions are “social systems whose participants avow belief in a supernatural agent or agents whose approval is to be sought” (Dennett, 2006, p. 9). Many would quarrel with this definition, but, as Dennett suggests, it does not really matter. He is going to discuss a range of topics in the neighborhood. What ties these topics together is that they all concern only natural phenomena and, Dennett advises, that they all should be studied by science.

Since there is a plethora of disparate items to discuss, I’ll begin by making some critical comments on five issues that arise from *Breaking the Spell*: (a) the conception of religion as a single topic, (b) the question of origins, (c) a matter of evidence, (d) meaning and values, and (e) some effects of religion. Then, at greater length, I’ll discuss three deeper issues that overlap with my work. In this section, I shall make use of some of my own writing to make some constructive (as well as critical) remarks on the three issues. Here are the three deeper issues: (A) anti-essentialism, (B) the intentional stance, and (C) the conception of persons. I shall write in the same informal style that Dennett uses to discuss these issues.

## 1. Brief Discussion of Five Issues

Let’s turn to the critical comments.

### 1.1 *The Conception of Religion*

Although he is well aware of the diversity of religions, Dennett conceives of religion as a single topic of investigation. However, it is very difficult to find a stable category of generic religion that can be studied scientifically. How do we determine what to include and what not to include as religion? There are many disparate kinds of candidates for religion—from Aztecs who cut out the hearts of captives in order to appease the sun-and war-god Huitzilopochtli to Jesuits who suffered torture and death for their missionary work in current-day Canada—not to mention Confucianism, or Cargo Cults, or even Communism. (Even the collecting and trading of baseball cards might carry the allure of religion for some.) I do not mean to suggest that this is news to Dennett; he mentions Cargo Cults and other distant practices. My point is just that all these practices remain dramatically different from each other, and there is no obvious basis on which to include or exclude a cultural practice in order to have a study of religion generically.

The question of whether religion was a single phenomenon was thoroughly discussed by 19th-century theologians. Schleiermacher argued that there is no whole that “religion in general” can signify. The term “religion” used as a general term piggybacks on particular religions, which are defined by communities or churches (Schleiermacher 1831, p. 32). Whether religions are natural phenomena or not, we need to distinguish one form of religion from another in order to understand them. If Schleiermacher is right, then there is no single topic of religion available for study by science.

## 1.2 *Origins of Religion*

Since Dennett treats religion as a single topic, I’ll say no more about the obstacles to doing so. Dennett (2006) says, “To understand religion today, as a natural phenomenon, we need to understand its origins” (p. 6). Extrapolating from evolutionary history, “we can surmise how folk religions emerged without conscious and deliberate design, just as languages emerged, by interdependent processes of biological and cultural evolution” (p. 114).

Supposing that Dennett is right about the origin of religion(s), why would this matter? Origins do not tell us what something is, or what function it serves. The postindustrial, digital environment that we live in bears little resemblance to the prehistorical environment in which religions evolved. Dennett (2006) himself says that religious ideas have been around longer than recorded history (p. 6). With language, which is needed in order for religions to exist, we can evolve culturally as well as biologically. Going back to the prehistoric origins of a kind of phenomenon tells us nothing about what it has evolved into—especially since, as Dennett says, we can transcend our genetic imperatives (p. 4).

Sharing a prehistoric phase need not connect us (or religions) today. Why dwell on origins? This is similar to explaining a complex cultural practice like art in terms of cave painting. It is no help to know anything about the earliest stages of art in order to understand art today (Danto, 1981). The same, I think, holds for religion.

## 1.3 *Evidence*

Dennett objects, “Given the way religious concepts and practices have been designed, the very behaviors that would be clear evidence of belief in God are also behaviors that would be clear evidence of (only) belief in belief in God” (2006, p. 223). This may be thought to vitiate evidence of belief in God inasmuch as almost everybody believes that people believe in God.



As an objection, this is a bridge too far. If it is an objection at all, it is an objection to evidence for belief in anything. Evidence that any language user who is not self-deceived had any belief in anything whatever would also be evidence that she had a belief in that belief. The evidence that I believe in tables, and not just in particles-arranged-tablewise, is also evidence that I have a belief in the belief in tables, and not just in particles-arranged-tablewise. If the locution “belief in tables,” say, is inappropriate, just change it to “belief that tables exist.” Then, the objection would be that any evidence for belief that tables exist would be evidence for the existence of belief that belief that tables exist. So the objection is empty, because it is true of all evidence for belief in anything.

### 1.4 *Meaning and Values*

According to Dennett (2013), Darwin’s main contribution was to “unite meaning with matter” (p. 201). However, it seems to me that Darwin achieved this union only by eviscerating the idea of meaning. “Meaning without a meaner,” along with “design without a designer,” is a clever slogan for a genuine accomplishment. But I think that the accomplishment is to give the *appearance* of meaning without a meaner, leaving it open whether there is any more to meaning than the appearance of meaning.

Dennett says that Darwin’s revolutionary idea does not destroy what we value, such as “ethics, art, culture, religion, religion, humor, and yes, even consciousness.” Rather, Dennett (2013) says that Darwin’s idea puts them on a better foundation, and “unites them gracefully with the rest of knowledge” (pp. 204, 203). It is difficult to give this claim credence when Dennett also says that Darwinian thinking “turns the whole traditional world upside down, challenging the top-down image of designs flowing from that genius of geniuses, the Intelligent Designer, and replacing it with the bubble-up image of mindless, motiveless cyclical processes churning out ever-more robust combinations until they start replicating on their own.” (p. 277). It would be a neat trick if Darwin turned the traditional world upside down without damage to what we value. It seems that Dennett wants to have it both ways.

### 1.5 *Some Effects of Religion*

According to Dennett, “The most pressing questions concern how we should deal with the excesses of religious upbringing and the recruitment of terrorists, but these can only be understood against a background of wider theories of religious conviction and practice” (2006, p. 307).

Let me make two comments on this claim. First, the “background of wider theories of religious conviction and practice notwithstanding,” there is too much variety in religious upbringings to suppose that there is a single phenomenon “religious upbringing” that we must “deal with.” How should we categorize religious upbringings? Should we categorize religious upbringings in terms of what they aim to produce—orthodox Muslims or adherents of ethical culture, say? Or, perhaps we should categorize them in terms of their methods—by example, say? I don’t see a way of drawing a line that distinguishes religious from nonreligious upbringings. Second, although there are religious terrorists, terrorists are not mainly products of religion. In the 20th century, Stalin’s and Hitler’s totally secular (not to say, nihilist) armies carried out more terror than any religious-backed groups. Terrorism and organized malice are spread around nicely; they are not confined to religion. Religious terrorists stand out, because the evil that they perpetrate is statistically unusual—compared to the evil perpetrated by Mao Tse Tung, Pol Pot, Shining Path in Peru, FARC (Armed Revolutionary Forces of Colombia), or the *NarcoTrafficantes* in Mexico, along with Hitler and Stalin. And that’s just in the 20th century, and not all of it.

Dennett continues, “Only when we can frame a comprehensive view of the many aspects of religion can we formulate defensible policies for how to respond to religions in the future” (2006, p. 310). Two more comments are in order: (a) It seems to me a pipe dream to suppose that “we can frame a comprehensive view of the many aspects of religion”; and (b) why do we need policies—defensible or not—to tell us “how to respond to religions in the future”? I think that the only policy we need is a secular state, with separation of Church and State (with continuing controversy about what it is permissible for churches and states to do). The current Supreme Court notwithstanding, people ought not to be allowed to violate the laws of the state in the name of religion any more than in the name of science, or in the name of nothing at all. In any case, I don’t see how we—whoever “we” are—could (or should) do more in terms of “defensible policies for how to respond to religions in the future.”

Perhaps more important is Dennett’s charge that religion promulgates “xenophobia, violence, the glorification of unreason, the spreading of patent falsehood” (Coyne, 2010). Maybe. But also in the Middle Ages, Christians established a hospital that became a center of medical training, and also “provided food for the hungry, cared for widows and orphans, and distributed alms to all who came in need” (Hart, 2009, p. 30). Quakers and Buddhists have been significant in antiwar activities, as have religious activists in the U.S. civil rights movement. It seems to me nigh well impossible to balance out the good and evil done by religious institutions.

Dennett has a sense of urgency about his project of breaking the taboo against scientific study of religion. He fears that “if we don’t subject religion to such scrutiny now, and work out together whatever revisions and reforms are called for, we will pass on a legacy of ever more toxic forms of religion to our descendants” (2006, p. 39). It seems to me that even tentatively lumping together all social systems whose adherents are committed to a supernatural agent in our scientific study of religion seems unlikely to provide a reasonable basis for a fear that we may “pass on a legacy of ever more toxic forms of religion.” It does not take science to tell us that any form of religion that opposes killing people is not as toxic as, say, the belief systems that supported the armies that systematically murdered, pillaged, and raped their way across Eastern Europe several times in World War II (Snyder, 2010).

## 2. Three Deeper Issues

Now I’ll turn to the three larger issues that overlap some of my own work—anti-essentialism, the intentional stance, and the conceptions of persons (Baker, 2007, pp. 34, 220; 1987, pp. 150–166; 2013, 127–128, 141–143). I’ll discuss these at greater length.

### 2.1 *Anti-essentialism*

Dennett (2013) points out that Darwin does not draw lines: there was no first mammal; and he uses the point to argue against essentialism (pp. 240–243). Dennett takes the fact that there was no first mammal, with which I fully agree, to imply that there are no essential properties. Since there are (obviously) mammals, Dennett takes the fact that there was no first mammal to unseat essentialism of any kind (pp. 241–242). Although I agree with Dennett that “we should quell our desire to draw lines” (p. 241), I do not think that anti-essentialism follows from the fact that there are no sharp lines in nature. There is no contradiction between the claim “there is no first mammal” and the claim “mammals essentially have the capacity to breastfeed their young.”

Let me explain: All that follows from the fact that there was no first mammal is that the biological kind *mammal* came into existence gradually. Well, that’s Darwin’s point. The dispositional capacity to breastfeed offspring was not instantaneously instantiated either; it came into existence gradually, along with mammals. Both the instantiation of the essential property of mammals and the existence of mammals came into existence together. The essential property of mammals is not essential to any particular mammal (e.g., a male). It is essential

to the biological kind.<sup>1</sup> Organisms of the kind *mammal* did not exist until the capacity to breastfeed the offspring (if that is an essential property of mammals) came into existence. What I think the point that there are no sharp lines in nature shows is that there is ontological vagueness, vagueness in reality. And, unlike Dennett, I do not believe that ontological vagueness leads to anti-essentialism (Baker, 2007, pp. 121–141).

On my view, every concrete thing that exists in the natural world is of some primary kind or other, and every primary kind has an essential property—the property that is required to be of the kind in question. This is true, I believe, of everything in existence in the natural world, including artifacts. Every mammal is of a primary kind, an essential property of which is the capacity to breastfeed the offspring. Every dialysis machine is of a primary kind, the essential property of which is to be intended to artificially remove waste from kidneys.

Both the primary kind and the individuals of the kind come into existence gradually. Not only is there no first mammal; there is no first moment of a particular mammal's existence. Consider the human organism. Sperm meets egg and fertilizes it. Fertilization is not instantaneous; it can take up to 20 hours (Ford, 2002, p. 55). Even then, a fertilized human egg may divide into twins, and hence is not yet a particular individual human organism, inasmuch as a single individual cannot become two individuals. (The logic of identity precludes any particular individual's becoming two; it is a necessary truth that if the individual splits, it is replaced by two successors, and the original individual goes out of existence.) The possibility of “twinning” ends when the fertilized egg is implanted in a uterus, but implantation of an egg in a uterus is not instantaneous either (Anscombe, 1985, pp. 111–112).

So, although I agree with Dennett that “we should quell our desire to draw lines,” I do not think that we should infer from this anything about anti-essentialism (Baker, 2007, pp. 121–141). If (pace Dennett) Darwinism does not require rejection of essentialism, a religious essentialist need not worry about this objection.

## 2.2 *The Intentional Stance*

Religion is ancient, perhaps even older than writing. Dennett (2006) has a just-so story about how the intentional stance evolved in response to problems in our environment, “Driven by the demands of a dangerous world, [the human mind]

---

1. To think that if *p* is an essential property of a kind, then every member of the kind must have *p* is, I suspect, to commit what we used to call “the fallacy of distribution.”

is deeply biased in favor of noticing the things that mattered most to the reproductive success of our ancestors" (p. 107). We need to recognize other creatures with minds, creatures whose behavior can be predicted by treating them as agents. False positives in identifying predators were not as dangerous as "false negatives" in which we failed to recognize a predator as an agent. So, natural selection opted for a hyperactive agent detection device, or HADD, which, on occasion, triggered an "agent there" response to something that was not an agent. This device in our ancestors' bag of tricks was the basis for what Dennett has made famous as "the intentional stance" (p. 108).

Nobody has to be taught that the world is full of agents who have beliefs and desires. We just experience other living beings as agents. Once our ancestors started talking, they learned to articulate the intentional stance: "A fox may be cunning, but a person who can flatter you by declaring that you are cunning as a fox has more tricks up his sleeve than the fox does, by a wide margin." It is quite plausible that the intentional stance evolved naturally, and that the activity of our hyperactive agent detectors lies at the root of human belief in gods—beings to whom we attribute agency (Dennett, 2006, p. 114). But it is something of a jump from hypothetical agent-detection devices that overreach to non-hypothetical religions.

This is a lovely story, according to which religion began and is sustained by the intentional stance. The intentional stance, evolved by natural selection, lies as the root of human belief in gods—beings to whom we attribute agency (Dennett, 2006, p. 114). Without foresight or intention, clever animals discovered "the Good Trick of adopting the intentional stance" (p. 109), which Dennett has discussed in detail in many places. But this same intentional stance brings into "focus a virtual world of imagination, populated by the agents that matter . . . to most to us" (p. 114).

Although I think that Dennett is entirely correct about the utility of the intentional stance, I also think that the intentional stance, on his conception, is unstable. Moreover, I think that if we stabilized it, the intentional stance could not be used to challenge religion. Indeed, the intentional stance would be indifferent to religion(s).

The instability that I see in Dennett's conception stems from taking the intentional stance to be a faculty with no ontological import: it allows us to treat moving bodies *as if* they were rational agents and to predict their behavior—as agents, those "rememberers and forgetters, thinkers and hoppers and villains and dupes and promise-breakers and threateners and allies and enemies" (Dennett, 2006, p. 111). It is rather like Ryle's inference-ticket. It tells us nothing about reality, only about what we are allowed to infer from what.

I can try to illustrate the instability of the intentional stance by “meme theory.” A meme is analogous to a gene, but instead of being a unit of biological significance that can be transmitted, a meme is a unit of cultural significance that can be transmitted. Examples of memes are fashions like wearing baseball caps backward or shaking hands (Dennett, 2006, p. 81). These fashions clearly require the intentional stance.

Now ask: Are memes (qua memes) discernible from the physical stance? Well, on the one hand, yes. Dennett (2006) speaks of memes as inhabiting the brain (p. 378). But if the brain is inhabited by memes, then there should be neural evidence of these items discernible from neuroscience. (I am not supposing that there will be any identification of memes with specific brain structures, but only that there must be some physical way that “the brain might store cultural information” (Dennett, 2006, p. 349).

Moreover, Dennett (2006) also holds that “the ultimate beneficiaries of religious adaptations are the memes themselves; but their proliferation (in competition with rival memes) depends on the ability to attract hosts *one way or another*” (p. 186). Presumably, the hosts that memes attract are (physical) brains, in which case there should be a physical difference between a brain that a certain meme inhabits, and a brain uninhabited by that meme. If so, then the meme (or its effect) should be discernible from the physical stance. (I confess that I find the supposition that memes benefit from the increased frequency of their occurrence to take the analogy with genes too far. It seems to me that to say that memes benefit from anything is just word play. Does singing “The Star-Spangled Banner” at football games benefit “The Star-Spangled Banner”? Really?)

So, from what I have quoted from Dennett, on the one hand, memes are discernible from the physical stance. But on the other hand, memes seem not to be discernible from the physical stance, but only from the intentional stance. For example, words, along with computer viruses, rituals, games (all mentioned by Dennett, 2006, p. 344) are memes whose individuation conditions are not physical, and hence are not detectable from the physical stance. Since the intentional stance allows memes to be both detectable and undetectable from the physical stance, we do not have a stable answer to the question, are memes discernible from the physical stance? Sometimes the intentional stance is taken to be metaphorical; at other times, it seems to yield a serious (ineliminable) description of reality, not just a convenient way to talk. The intentional stance seems to wobble. Hence, the instability. (For an extended argument taking issue with Dennett’s view on this matter, see Baker, 1987, pp. 149–163.)

However, on a different conception (my own), an analogue to the intentional stance is stable. On my conception, the analogue to the intentional stance is not

just a useful strategy for prediction, but it gives us a generally reliable purchase on reality—genuine reality, as real as electrons. Even after “Real Patterns” (Dennett, 1991), I am unsure about Dennett’s *ontological* view of the intentional stance; but my own view is that (an analogue of) the intentional stance reveals parts of reality that are not revealed in any other way (say, by the physical stance). From the physical stance, I can detect that moving bodies are entering the room; but physics, I strongly suspect, will never allow me to detect that the Queen herself is entering the room.

It seems to me patently false to suppose that all of reality can ever be detected from the physical stance. It would be inaccurate to describe my waving when I see you as my hand’s rising on seeing a moving body. I think that Dennett would agree; but if so, why not accord ontological status to the (veridical) deliverances of the intentional stance? Why reserve ontological significance for what is discerned from the physical stance alone (Dennett, 1978, pp. 27, 285)? In any case, if I am right, then the intentional stance is not merely a strategy, or a handy way to talk, or a way to avoid the messy details of physics; it is rather a window onto reality itself.

My revision of the intentional stance would allow Dennett’s evolutionary story to remain in place. Our “agent-detecting,” though not infallible, is generally reliable. When I hear my husband coming home, I detect him, and surely, he—my dear husband, Tom—really exists. On most occasions, I detect genuinely real entities (like my husband) from the intentional stance; however, on other occasions I “detect” (or just seem to detect) nonexistent entities (like burglars in the house). However, the fact that the intentional stance is fallible in itself gives no reason to doubt the veracity of religious beliefs. I see no non-question-begging way to rule out accurate detection of religious entities from the intentional stance. Since the intentional stance is generally trustworthy, mere fallibility cannot be used against detection of religious entities. Every means that we have of detecting things is fallible. So, even if Dennett is right that the fallible intentional stance sustains religion, I do not think that the intentional stance can be used to cast doubt on religion.

### 2.3 *The Conception of Persons*

Dennett’s view of persons by itself seems to stack the deck against the three great monotheisms. The great religions all construe human persons as morally significant beings. For Dennett, persons are the result of our ancestors’ learning “Good Tricks” or adding to their “bag of tricks.” The “human mind is something of a bag of tricks, cobbled together over the eons by the foresightless process of evolution by natural selection” (2006, p. 107).



Dennett construes persons as assortments of mind-tools and other tools, perhaps as bags of tricks.<sup>2</sup> The user of the tools seems to disappear; we seem to be tools “all the way down.” Dennett does not actually call us contemporary persons bags of tricks (as far as I can tell).<sup>3</sup> But he does argue that what we think of as consciousness is an illusion; “consciousness is a bunch of tricks in the brain” (Dennett, 2009). A bunch of tricks seems unlikely to have moral significance or to be deserving of being treated with any dignity.

One may object: The tool-talk is just a metaphor, not to be taken literally. Then, I ask: What does Dennett take literally? I do not see any other candidates for what persons are in Dennett’s writings—at least recent writings—other than as *assortments of disparate items*, and it is that feature that I think makes Dennett’s view of persons incompatible with Judaism, Christianity, and Islam. Although I do not expect Dennett to care whether his view of persons is incompatible with the great monotheisms, I do expect him to care about whether persons in their own right deserve to be treated with dignity.

In the first place, on Dennett’s view, a person is not characterized by any kind of unity. Each person is just a product of haphazard accumulation of tools and traits over the course of evolution, genetic and cultural. I agree with Dennett that there is no Cartesian Theater, a place in the brain where everything comes together, no self other than the whole embodied person. Persons are “embodied and embedded,” as the slogan goes (Baker, 2009). But I do not think that it follows that we are just bunches of tricks. And if we were, why should be accorded dignity, or any more value, or any more moral significance than the items in my toolbox in the garage? But if there’s no Cartesian Theatre, and we have evolved in the way that Dennett describes, wherein lies our moral significance as individuals?

Assuming that bags of tools lack the unity required for persons to be morally significant, I have a suggestion to offer on Dennett’s behalf: one may agree with Dennett that the organisms that evolved into beings that constitute us are no more than bags of tricks. Fine. The human organisms resulting from natural selection are bags of tricks, but we—human persons—are not just the resulting organisms. What sets us part is that we have language, and all that language entails. Our languages have devices that support a robust first-person perspective—a capacity to conceive of ourselves as ourselves in the first person. And this capacity increases our causal powers to the extent that we are more than the organisms

---

2. In *Brainstorms*, Dennett (1978) thought of persons entirely in terms of the intentional stance—far too instrumentalistically for me.

3. In 1973, Dennett referred to a personal stance as a moral stance that presupposes the intentional stance, but I do not think that he has discussed this in recent years (1973, pp. 166–167).



that constitute us (Baker, 2013). The robust first-person perspective distinguishes us persons from every other kind of thing in the universe (Baker, 2013).

The person is constituted by the organism (the bag of tools), but is not identical to any organism. The person has a first-person perspective essentially, and it is the first-person perspective that provides the unity that gives persons moral value. A robust first-person perspective is a person-level dispositional property (Baker, 2013). To put it metaphorically, the bag of tools initially has loops around the top through which a string lies open. That's the human organism at a fetal stage—evolutionarily, morphologically, and anatomically continuous with nonhuman organisms. When the human organism develops to the point of supporting intentionality and consciousness and a second-order capacity for self-consciousness, the string at the top of the bag of tricks is tied. Then, a unified person emerges—a person constituted by, but not identical to, an organism. So, a human *organism* is the bag of tools that Dennett describes. But the *person* is more: the person is unified by the first-person perspective—a person-level dispositional property. Persons have moral significance that does not depend on anyone else's third-personal intentional stance. They deserve to be treated with dignity, not because others assume an intentional stance toward them, but because of what they really are.<sup>4</sup>

The point about the moral significance and deserved dignity of persons seems to me to be a good reason not to suppose that science is the exclusive arbiter of genuine reality (i.e., of what entities, properties, and kinds exist irreducibly and ineliminably). Since science does not lead to a conception of people as unified individuals, science will not find people to deserve to be treated with dignity because of what they are. I think that this leads to a case of “one philosopher's modus ponens is another's modus tollens.” From “if science is the exclusive arbiter of all genuine reality, then it is not in virtue of what they are that people deserve to be treated with dignity,” I would infer by modus tollens that science is not the exclusive arbiter of all genuine reality. But I suspect that Dennett—while actually treating people with dignity—would infer by modus ponens that it is not in virtue of what they are that people deserve to be treated with dignity.

### 3. Conclusion

The idea of religion as a single, natural phenomenon arose in the Enlightenment. Since then, religion has not been treated with deference among the learned.

---

4. Although this view does not spring from science, I do not believe that it contravenes any extant science (Baker, 2007, pp. 67–93).

The 19th century had a number of great German theologians (Friedrich Schleiermacher, David Friedrich Strauss, Ernst Troeltsch, and Johannes Weiss) who used scientific historical methods to question the authenticity of much of the Bible.<sup>5</sup> Viewing the Bible as texts with human (not supernatural) origins, their new-critical approaches “left the old Biblical certainties in tatters” (Wilson, 1999, p. 343). In light of the sustained criticism from scholars since the 19th century—not to mention Hume’s treatment of religion in the 18th century—it seems somewhat hollow to suppose that religion has escaped scientific scrutiny.

Nevertheless, I agree with Dennett that the general population—with its climate-change deniers, natural-selection deniers, religious theme parks, and so on—could benefit from having more respect for science. So, my discussion notwithstanding, I think that *Breaking the Spell* may help foster that respect.

## Acknowledgments

Many thanks to Katherine Sonderegger for help on this paper.

## Works Cited

- Anscombe, G. E. M. (1985). “Were you a zygote?” In A. Phillips Griffiths (Ed.), *Philosophy and practice* (pp. 111–115). Cambridge, England: Cambridge University Press.
- Baker, L. R. (1987). *Saving belief: A critique of physicalism*. Princeton, NJ: Princeton University Press.
- Baker, L. R. (2007). *The metaphysics of everyday life*. Cambridge, England: Cambridge University Press.
- Baker, L. R. (2009). Persons and the extended-mind thesis. *Zygon: Journal of Religion and Science*, 44, 642–658.
- Baker, L. R. (2013). *Naturalism and the first-person perspective*. New York, NY: Oxford University Press.
- Coyne, J. (2010). Kitcher versus Dennett: Is new atheism counterproductive? [Web blog post]. Retrieved from <http://whyevolutionistrue.wordpress.com/2010/10/07/kitcher-versus-dennett-is-new-atheism-counterproductive-2/>
- Danto, A. C. (1981). *The transfiguration of the commonplace*. Cambridge, MA: Harvard University Press.

---

5. As Albert Schweitzer (1910) put it at the beginning of the 20th century, “The critical study of the life of Jesus has been for theology a school of honesty. The world had never seen before, and will never see again, a struggle for truth so full of pain and renunciation as that of which the lives of Jesus of the last hundred years contain the cryptic record.” (p. 5)

- Dennett, D. C. (1973). Mechanism and responsibility. In Ted Honderich (Ed.), *Essays on freedom of action* (pp. 157–184). London: Routledge and Kegan Paul.
- Dennett, D. C. (1978). *Brainstorms: Philosophical essays on mind and reality*. Cambridge MA: MIT Press.
- Dennett, D. C. (1991). Real patterns. *Journal of Philosophy*, 88, 27–51.
- Dennett, D. C. (2006). *Breaking the spell: Religion as a natural phenomenon*. New York, NY: Penguin.
- Dennett, D. C. (2009). Daniel Dennett explains consciousness and free will. Retrieved from <http://bigthink.com/videos/daniel-dennett-explains-consciousness-and-free-will>
- Dennett, D. C. (2013). *Intuition pumps and other tools for thinking*. New York, NY: Norton.
- Ford, N. M. (2002). *The prenatal person*. Malden MA: Blackwell Publishing.
- Hart, D. B. (2009). *Atheist delusions: The Christian revolution and its fashionable enemies*. New Haven, CT: Yale University Press.
- Schleiermacher, F. (1831). *The Christian faith*. (2nd ed.). Edinburgh, Scotland: T&T Clark.
- Schweitzer, A. (1910). *The quest for the historical Jesus*. London. Adam and Charles Black.
- Snyder, T. (2010). *Bloodlands: Europe between Hitler and Stalin*. New York, NY: Basic Books.
- Wilson, A. N. (1999). *God's funeral*. New York, NY: Norton.

# 11.2 REFLECTIONS ON LYNNE RUDDER BAKER

Daniel C. Dennett

I don't think I could name a philosopher I'd rather see write an essay on *Breaking the Spell* than Lynne Rudder Baker. She's is a serious philosopher of mind and metaphysics, a longtime<sup>1</sup> radical critic of the intentional stance *and* a Christian apologist, who gave one of the Gifford Lectures on Natural Theology in Glasgow, in 2001. And unlike some other distinguished religious philosophers who shall remain nameless, she somehow summons the discipline to tell the truth about my inflammatory claims. Not for her the Trumplike lies (so transparent and readily rebutted that I often call them "faith fibbing" just to embarrass their perpetrators). One of her signature projects is the defense of a "Constitution" view of persons: Yes, we human beings are animals, but we are also (constituted as) persons, with unique powers from which it follows that we alone are truly moral agents. She is thus a fearless exponent of human exceptionalism, but no creationist. She thinks we can have science and Christianity in harmony:

A proponent of the Constitution View need not postulate any gap in the animal kingdom between human and non-human animals that is invisible to biologists. Nor need the Constitutionalist deny natural selection. She may insist that theists can and should give science its due. Circling the wagons against the onslaught of modern science is hopeless; it just breeds a kind of defiant brittleness and alienates theists from the world that they cannot avoid living in (Baker, 2008)

---

1. She was a vigorous participant in a seminar I gave at Pittsburgh, in 1977, going through most of the material in *Brainstorms* (1981), and should have had several footnotes in that book acknowledging improvements she eked out of me.

So one might think that she would support my version of human exceptionalism, grounded in Darwinian natural selection and brought to life by the vast explosion of comprehension and insight engendered in us by *cultural* evolution, which only the human species enjoys in transformative measure. But no, she is at pains to distance herself from my bottom-up, naturalistic *elevation* of our species above the rest, which by her lights is an unworthy substitute for a more traditional version that descends from the myths of Western Civilization, and of Christianity in particular. But aside from these different ways of charting sources (themes that *might* be cast in aesthetic, rather than scientific, terms), our views on the nature of persons are very similar. In such close quarters the temptation to exaggerate the differences is often irresistible, but not to Baker. She is almost always accurate in saying what my views are, even though on occasion she seems to underestimate the ability of my position to rebut her objections.

For instance, before turning to her discussion of “deeper issues” she lays out five “critical comments” in ten swift paragraphs, and I was surprised by these, not because they were unfamiliar to me, but for the opposite reason: these are all shopworn crowd-pleasers, a sort of How to Challenge an Atheist catechism of zingers, and her cursory treatment of them suggested to me that perhaps she felt a political duty to her co-religionists to get them into the record, so to speak, rather than thinking of them as serious obstacles to my case. Acting on that assumption, I will be even more cursory in rejecting them:

1. *The Conception of Religion*. Yes, there are many hugely different forms of religion, so it is hard to generalize. But not impossible. As I say in *Breaking the Spell*, I do discuss the outliers, some of which we may consider honorary religions, just to get them into the investigation. Was *Archaeopteryx* a bird? Yes, “by definition”; what about its ancestors? Let’s not argue; when in doubt, include them. Baker says that “there is no obvious basis on which to include or exclude a cultural practice” (in the set of religions), and we Darwinians agree there is no *obvious* basis, but we do not consider that an objection, for the reason enunciated below.
2. *Origins of Religions*. “Origins do not tell us what something is, or what function it serves.” On the contrary, origins tell us a great deal—but not everything—about what something is, and what functions it serves or has served, and much of this is not obvious. “Going back to the prehistoric origins of a kind of phenomenon tells us nothing [*sic*] about what it has evolved into—especially since, as Dennett says, we can transcend our genetic imperatives.” Not only does it tell us a great deal; it typically explains why we might want to include a phenomenon in a set—a lineage—of vastly varying things (see 1.1). The discovery of well-nigh invisible differences and similarities is often the key to revolutionary changes in our knowledge of what something is.

Armchair essentialism, in contrast, sticking—as it must—to intuitions about what is obvious, is systematically out of synch with science, forever defending the obsolete from the timeless perspective of “eternal” essence. (And, when we do transcend our *genetic* imperatives, it is thanks to the historical process of *memetic* selection, so here, too, d’Arcy Thompson gets it right: everything is the way it is because it got that way.)

3. *Evidence.* I argue that more people believe in belief in God than believe in God. How do I know? Because—presumably—those who believe in God *also* believe in belief in God (that is: they believe that belief in God is a *very good thing*), and added to that population, however large it is, are those who wish they could believe in God (they believe in belief in God without, alas, believing in God). Here Baker does mistake my position, treating belief in belief in God as trivial, “Almost everybody believes that people believe in God.” So she rebuts a phantom position. I grant that “belief in belief (in God)” *could* be interpreted to mean “belief that there is belief (in God)” —after all, belief in Satan would not *normally* be understood to mean belief that Satan is good, but in the context of my argument it is manifest that “belief in belief” takes the approval reading. When somebody says “I believe in birth control,” they are not taken to mean that they think birth control exists.
4. *Meaning and values.* “It would be a neat trick if Darwin turned the traditional world upside down without damage to what we value.” Indeed, it would, and it was, and that’s why Darwin’s dangerous idea is the best idea ever. Baker’s “critical comment” is actually just a candid expression of utter disbelief in the main (argued) claim of *Darwin’s Dangerous Idea* (1996), not an objection to it.
5. *Some Effects of Religion.* In this comment, Baker does her version of *tu quoque*, conjuring up the shades of Hitler, Stalin, Mao Tse Tung, Pol Pot, et al., to which the appropriate answer is, as always, Yes, those were all atrocities committed by atheists, and now we’re going to focus on the atrocities committed in the name of religion, atrocities that have tended to get a free ride because of the traditional mantle of respect for religious belief. She asks:

Why do we need policies—defensible or not—to tell us ‘how to respond to religions in the future’? I think that the only policy we need is a secular state, with separation of Church and State (with continuing controversy about what it is permissible for churches and states to do).

Exactly, except that she seems to be suggesting that the “continuing controversy” is a smallish matter of differing opinion that will be best addressed if we continue to maintain the default respect that keeps the empirical details fuzzy and the embarrassing questions unasked.

That is all that needs saying about the critical comments, in my opinion: she has raised her battle flag and I have raised mine in salute. Then she turns to three deeper issues, concerning my anti-essentialism, the intentional stance, and the conception of persons.

## 1. Anti-essentialism

When Baker agrees with me that there was no first mammal and that we should quell our desire to draw lines, but goes on to say that this does not lead her to deny that there are essences, my first thought is that she maybe holds a view like Douglas Hofstadter's. When I queried him about his use of "essences" in the title of his recently co-authored book, *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking* (with Emmanuel Sander, 2013) he responded:

I don't react negatively or fearfully when I hear the word "essence", because in my mind the word has not been tainted by a wearying set of arcane debates in philosophical circles. For me, it's just an informal, everyday word, not a technical term. I doubt that you and I have any disagreement about what the word "essence" means when it's used informally, as a synonym for "gist", "crux", "core", etc. And that's how it's used in the book. (personal correspondence, January 3, 2014)

He is right that I have no disagreement with him regarding his (familiar, non-technical) sense of the term. I worry, however, that the excellent insights he and his co-author reap from their celebration of essences in *this* sense will lend false respectability to the philosophers' "arcane debates" when they use the term—and more recently, when they use thinly veiled substitutes (euphemisms, in effect), now that its credentials have been challenged.<sup>2</sup>

Baker's essentialism is not, I have learned, Hofstadter's innocent reliance on an informal colloquialism, but a full-throated, technical defense of a kind of essentialism that really does posit essences as ineliminable features of our *practical* metaphysics. This, too, might well find favor with me, given my insistence on taking the ontology of the affordances of everyday life more seriously than the ontology of any other metaphysics (see Kukla, chapter 1.1, this volume, and my reflections, chapter 1.2). Might there be an irenic merger between my affordances and Baker's essences? Perhaps, but I am dubious about her "primary

---

2. These paragraphs have been drawn with minor revisions, from my "Darwin and the Overdue Demise of Essentialism," (2017b)

kinds” and a threatened population explosion: is every computer of a single primary kind, or are digital and analog computers of different primary kinds, and if the latter, is a MacBook Pro with a Retina display of one primary kind, while a MacBook Pro with a different display of a different primary kind? I cannot claim to have mastered the details of Baker’s version of essentialism, and perhaps elsewhere she provides compelling reasons for why we *need* such a notion, but I am pushed into skepticism from the outset by some of its indirect fruits: as I noted in my earlier comments, it surprises me that she underestimates the power of natural selection to provide key ingredients in genuine answers—right or wrong—to her deeper philosophical questions. It strikes her as just obvious that the world of meaning and value could get no illumination from the world of evolutionary biology—so obvious that no supporting argument or analysis is called for. Why not? Would it be because evolutionary biology cannot get to the *essences* of these phenomena? She is far from alone among philosophers in this regard. Many who would agree with her advice “no need to deny natural selection” never even *consider* the prospect that natural selection has implications for their pet topics (meaning, reality, truth, knowledge, ethics, free will, . . .). Expressing my skepticism here is of course no argument against her essentialism, but I have presented arguments drawing out those implications, lined up in *Darwin’s Dangerous Idea* (1996), *Freedom Evolves* (2003), *Breaking the Spell* (2006), *Intuition Pumps* (2013), and, now, in *From Bacteria to Bach and Back*, (2017a) and I don’t think I need to repeat them here. That is, I know that some philosophers who have not read these books have heard from others that they are just popularizations of Darwinian themes, maybe “full of fascinating ideas” but not of ideas and arguments addressed directly against some of the most entrenched prejudices of academic philosophy. My arguments may all miss their targets, but we won’t know until some of those who endorse those targets, or take them for granted, confront those arguments.

## 2. The Intentional Stance

“Although I think that Dennett is entirely correct about the utility of the intentional stance, I also think that the intentional stance, on his conception, is unstable.” (p. 338) Her argument for this instability invokes my meme theory, and would kill two birds with one stone if successful. After quoting me on whether memes are discernible only from the intentional stance or also from the physical stance, she says:

So, from what I have quoted from Dennett, on the one hand, memes are discernible from the physical stance. But on the other hand, memes seem



not to be discernible from the physical stance, but only from the intentional stance. For example words, along with computer viruses, rituals, games (all mentioned by Dennett 2006, p. 344) are memes whose individuation conditions are not physical, and hence are not detectable from the physical stance. Since the intentional stance allows memes to be both detectable and undetectable from the physical stance, we do not have a stable answer to the question, Are memes discernible from the physical stance? Sometimes the intentional stance is taken to be metaphorical; at other times, it seems to yield a serious (ineliminable) description of reality, not just a convenient way to talk. The intentional stance seems to wobble. Hence, the instability.

Something has gone awry here. I am unclear what Baker thinks I mean about detectability, and my recent rereading of her *Saving Belief* (1987, pp. 149–163) did not dissipate my confusion. Obviously, the English word “cat” has (physical) tokens that share no distinguishing physical properties (think of spoken, whispered, hand-written and chiseled-in-stone tokens, for instance) in virtue of which they are all tokens of “cat” and yet it is easy enough to detect most tokens of the word “cat” by their physical properties. (We can hear, we can read.) Right now, say to yourself silently: “cat, cat, cat.” You just created three physical tokens of the English word “cat” and they had no acoustic properties in common with tokens spoken aloud (they were soundless *representations* of sounds, like the grooves on a vinyl record); indeed, we don’t yet know what their distinguishing physical features are, and cannot identify them (yet) in your brain, but we’re getting closer. That is, neuroscience is homing in systematically on how physical events in our brains can be representations of words (spoken, written, merely thought), or of houses, mountains, violins, . . . and in no case is it by being replicas, or by sharing (except coincidentally) physical features with what they represent. You couldn’t conduct this research without having the intentional stance because you wouldn’t be able to specify what you were looking for! That’s the sense in which memes, etc., are stance-dependent. I don’t see any instability. If there is instability, there must also be instability for the bland view that reading and aural linguistic communication are everyday physical phenomena, but you can’t identify what is being communicated just from the physical properties of the tokens, if you don’t know there’s a language involved, and which language it is. (For more on this, see “Two Black Boxes” in *Darwin’s Dangerous Idea*, 1996)

I can’t resist responding to one of Baker’s rhetorical questions about memes: “Does singing ‘The Star-Spangled Banner’ at football games benefit ‘The Star-Spangled Banner’? Really?” Really. In exactly the same way broadcasting your cold virus in a sneeze benefits the virus, not you. Can something *benefit* a

virus? It's just a complex unliving, unloving macromolecule! Yes, it can, in the uncontroversial Darwinian sense of enhancing its replication. (The campaigns against HIV, Ebola, and the Zika virus are *best understood* as concerted efforts to deny that benefit to these lifeless, mindless agents.) Baker sees no conflict between natural selection and her Constitution View, but her incredulity here says she should look again.

### 3. The Conception of Persons

"A bunch of tricks seems unlikely to have moral significance or to be deserving of being treated with any dignity." Why? Because dignity must be a trickle-down blessing, not a bubble-up achievement? That seems at first to be the suggestion, but Baker goes on to grant that a human *organism* is, as I suggest, a "bag of tools"; it's the *person* who has moral significance, in virtue of having language, and, thanks to language, "the first-person perspective" which "increases our causal powers to the extent that we are more than the organisms that constitute us." I agree with everything in this claim except the last phrase, which I guess I don't understand. This is in fact the view I have defended since "Conditions of Personhood" (1976). It is language that distinguishes human consciousness from mere animal consciousness, and that gives us the vastly enhanced powers of cognition and comprehension and self-control that *qualify* us as moral agents. And as for "constitution," I stressed that "person" is far from synonymous with "human being" and is a moral—or as Locke said, a forensic—concept, with a strikingly different role to play in human life. We are human beings *and also* persons. Is that what "constitution" means? If not, I don't understand.

Baker's only comment on my version is in footnote 2: "In *Brainstorms* [where "Conditions of Personhood" is reprinted], Dennett thought of persons entirely in terms of the intentional stance—far too instrumentalistically for me." Casting about for an interpretation of this, I come up with this:

According to me, we persons have framed a conception of ourselves as morally competent agents—persons, with responsibility and dignity—and we view this as a status we grow into, as our moral education achieves its ends, and can grow out of, if we lapse into senile dementia, for instance. Infants are potential persons; kittens and puppies are not, but might have been. Someday robots may be persons. At any one time, there is something approaching a consensus about which competences are needed to qualify one as a person in good standing, and there are reasons (good, ultimately consequentialist reasons) why we distinguish such a subset of human beings for special status: in a nutshell, it is a robust, practical

societal arrangement for maximizing trust, security, freedom, and tolerance. There is no criterion for being a *real* person beyond the current consensus, which is enough to stabilize and justify our treatment of each other, including both praise and blame, reward and punishment.

That is “instrumentalistic” in the sense that this instrument that society has constructed is itself the arbiter, not merely a reliable indicator, of the “essence” of personhood. To think otherwise would be like wondering whether the rules of baseball manage to define what a home run *really* is. (I would rather avoid the word “essence” altogether, but I recognize that some might find it useful to speak of *nominal* essences, of home runs, checkmate, personhood, and the like.) There is no Prime Mammal and there is no First Moment of personhood, when, miraculously, a person is constituted out of a human organism. I am not sure that Baker disagrees with this, but if so, I don’t understand her allergy to instrumentalism in this context.

Baker has given me a lot to think about, and I am somewhat uncomfortable to discover that my responses seem altogether *too obvious*. According to my analysis, she has misjudged my (outrageous) views as needing no more than an amused nudge to topple over. (“Really?”) And I worry I have suffered the mirror-image failure of imagination in my reply. I may well be missing some big points.

## Works Cited

- Baker, L. (2008). Our Place in Nature: Material Persons and Theism. Presented at a conference on Philosophy of Religion and Theology, Suffolk County Community College (Ammerman Campus), May 3, 2008.
- Baker, L. R. (1987). *Saving belief: A critique of physicalism*. Princeton, NJ: Princeton University Press.
- Dennett, D. C. (1981). *Brainstorms: Philosophical essays on mind and psychology*. Cambridge, MA: MIT Press.
- Dennett, D. C. (1976). Conditions of personhood. In A. O. Rorty (Ed.), *The identities of persons* (pp. 175–196). Berkeley: University of California Press.
- Dennett, D. C. (1996). *Darwin’s dangerous idea: Evolution and the meanings of life*. New York, NY: Simon and Schuster.
- Dennett D. C. (2003). *Freedom evolves*. New York, NY: Viking.
- Dennett, D. C. (2006). *Breaking the spell: Religion as a natural phenomenon*. New York; Penguin.
- Dennett, D. C. (2013). *Intuition pumps and other tools for thinking*. New York, NY: Norton.

- Dennett, D. C. (2017a). *From bacteria to Bach and back*. New York, NY: W. W. Norton and Company.
- Dennett, D. C. (2017b). Darwin and the overdue demise of essentialism. In D. L. Smith (Ed.), *How biology shapes philosophy*. Cambridge, England: Cambridge University Press: 9–22.
- Hofstadter, D., & Sander, E. (2013). *Surfaces and essences: Analogy as the fuel and fire of thinking*. New York, NY: Basic Books.



## INDEX

- 720-Flip, 207, 210–215
- abstraction, 96, 182, 189, 237, 239, 240, 243, 278
- ACT-R, 72–77
- agency, xix, xxvii, xxxi, 50–52, 296–298, 303–304, 309n8, 320–321, 328–330, 338; agency, joint, 311, 317
- Ainslie, George, 299, 303
- algorithms, Darwinian, 255
- algorithms, greedy, 257–258, 263, 290
- amnesia, 68–69, 80–81
- Amodio, David, xxx, 275–277, 300
- anarchism, 318, 330
- anti-essentialism, xxxii, 336–337, 348–349
- anticipation. *See* expectations; prediction
- artificial Intelligence, xxviii, 32, 44, 204
- associations, xv, xvii, 70, 108, 120, 210, 227–228, 230n11, 232–233, 235, 238, 242, 251–252, 276–280, 299–300, 320n15, 321
- atheism, 331, 346–347
- attention, xiv, xxii–xxiii, 8, 19, 23–25, 64, 73, 77, 79, 81, 158, 183–184, 187–188, 190–192, 216, 276, 295, 297, 309, 311, 321
- attitudes, implicit, 8, 15, 19, 274, 276–281, 282–283, 320n15
- attitudes, propositional, 4–7, 9, 17, 22, 26, 33, 43–45, 47
- Bain, Alexander, 228, 236
- Barrett, Lisa Feldman, xxii, 302, 310
- Barsalou, Larry, xxx, 278
- Bartlett, Frederic, 64, 74, 315
- Bayesian processes, xxiv, 72, 75, 92–93, 202, 206–210, 212–216, 219–221, 306, 307n6, 311, 319, 329
- belief-in-belief, 333, 347
- biases, xxxii, 15, 65, 275–276, 280–281, 305, 311, 313n10, 320n15, 321, 328–329, 338
- BitTorrent protocol, 316
- blindsight, 149
- Block, Ned, 155n13, 156n15, 159, 184
- Boghossian, Paul, 21, 28
- Brainese, 179
- Bratman, Michael, xxxii, 296–299, 304, 309, 313, 316, 319
- brute force, 111, 118, 262
- CADBLIND, 204–205
- Cartesianism, xv–xvi, 59, 143n10, 144, 149, 188, 293, 321; Cartesian Theater, xxi, 143, 186, 341, 282
- category-learning, 76–77

- centers-of-gravity, 5–6, 12, 34, 165  
 Central Dogma, 291–292  
 change blindness, 157, 183, 200–201  
 Chomsky hierarchy, 251  
 Chomsky, Noam, 99–100, 104, 111, 114, 115–116, 118–119  
 coffee, xx, 34, 209, 295, 297, 298, 299, 300, 301, 306–307, 308, 309, 310, 315  
 cognitive prosthetics, 306–308, 320  
 color-phi, 141, 157, 185–186, 187–188, 193, 200  
 competence, nonconceptual, 44, 45n2, 46, 49  
 compatibilism, xxvii, xxviii–xxx, 47  
 computation, xvii–xviii, xxi–xxii, 42, 62, 65, 72–78, 82, 95, 98–99, 178, 191, 197, 299, 301, 304, 316; computational competition, xxii, 142, 145, 152, 169, 251; Computational intractability, 44–45. *See* Association; Bayesian processes; expectations  
 concepts, xx, xxx, 11–12, 18, 26, 29, 36–37, 39–40, 43–44, 47–48, 57, 59–60, 81, 97, 113, 129, 190, 237, 244, 277, 278–279, 299, 301, 303, 312; phenomenal concepts, 134n2  
 conformity, xxix, 305, 313n10  
 connectionism, 102, 103, 114–115  
 consciousness, AIR Theory, 187–191  
 consciousness, fame in the brain, xxi–xxii, 186–187, 189  
 consciousness, illusionism, xix, xxii–xxiv, 59, 128, 142–143, 171, 174, 180, 184–185, 193, 220, 304, 341  
 consciousness, Multiple Drafts Model, xxi, 142, 144, 186  
 consciousness, Orwellian and Stalinesque, 141–143, 165–166, 185, 188, 200  
 consciousness, phenomenal, 159, 186, 191, 198; phenomenal concepts, 134n2; phenomenal properties, 197, 198, 207. *See* qualia  
 consistency, interpersonal, 309  
 constitutive misrecognition, 12n2, 20n7  
 constructional idioms, 107–108  
 content, mental, xvi–xviii, xxi, xxiv, xxx, 9, 20, 26, 46–47, 50, 53, 59n2, 66, 78–79, 81–82, 92, 152–153, 154, 156–157, 161, 165, 169, 174–175, 189, 191, 196, 255, 260, 276, 312  
 coordination, xxxi, 4, 7–9, 11–12, 14–15, 17, 19, 27–28, 32, 296, 310–312  
 cooties, 13n3, 34  
 coping strategies, xiv, 4, 7, 9, 11, 13–14, 15, 17, 19–21, 26n10, 29, 33  
 counterfactual thinking, 71–72, 240n29  
 crows, xxviii, 238–239  
 Cubism, 176  
 cuteness, xxiv, 202, 206, 212–216, 219–220  
 Darwin, Charles, 202, 237, 260  
 Darwinian thinking, vii, xi, xxiv–xxvi, xxx, 129, 225–226, 227, 250, 256, 258, 259–260, 261, 262, 263, 267, 273, 290–291, 306, 334, 336, 337, 346–347, 349–351; Darwinian compatibilism, xxviii; Darwinian humanism, vii, xi, xxvii, xxxi–xxxii; neuronal Darwinism, 257  
 data structures, xvii, 96–99, 128, 197  
 Dawkins, Richard, 254, 259, 290, 293, 331  
 Davidson, Donald, 24–26, 35, 58  
 default network, 70–72, 308n7  
 deflationism, xxiv, 26–28, 48, 186, 188, 189, 268  
 Diffusionism, 274

- dignity, xxxii, 341–342, 351–352  
discernment, 10–11  
double-consciousness, xxxi, 312n9  
double-transduction, 197, 202–204, 208, 216, 221  
Dretske, Fred, 46, 143n10, 151n11  
duck-rabbit, 173, 177
- echo chambers, epistemic, 316–317  
“elbow room,” 296, 317–318, 330  
embodiment, xiv, xxiv, xxviii, 4, 7–10, 13–14, 16, 20, 21, 26n10, 29, 278, 317, 341  
essentialism, xxxii, 59, 130, 283, 336–337, 342, 347, 348–349, 352  
evolution, culture, xxix–xxx, xxxii, 259, 261–262, 273, 317, 333, 346  
evolution, Darwinian, 233, 236, 244, 250, 251, 255, 256, 317; Darwinian spaces, 292; differential reproduction, xxv, 265, 266; natural selection, xv, xxiv–xxv, xxviii, 19, 48, 52–54, 59, 226, 227, 251, 256–257, 259, 266, 290, 293, 338, 340–341, 345–346, 349, 351; Universal Darwinism, 257  
evolution, gradualism, xi, xxv, xxxi, 129, 225, 236, 336, 337  
evolution, Lamarckian, 236, 245, 273, 291–292, 293  
expectations, xiii, xiv, xvii, 8–11, 13, 15, 16, 19, 63–64, 67, 70–72, 78, 83, 84, 206–207, 209–215, 236, 295n1, 302–303, 304–305, 309, 310–311, 316, 317, 320, 328. *See* prediction  
experience, subjective, xx–xxii, xxv–xxvi, 203  
exploration vs. exploitation, 267
- fashion, 281, 339  
filling-in, perceptual, 180–182, 184–185, 187, 193, 198, 200–201
- Fodor, Jerry, 41–42, 46, 78  
folk-psychology, xiii, xiv–xvi, 5, 6, 8, 20, 33–34, 36–37, 39, 40–43, 45, 48, 49–54, 141, 142n9, 152–153, 154, 156–157, 160, 169, 215, 234  
foreknowledge, xxxi, 298–300, 304, 314  
FPO (first-person operationalism), 142–148, 151–153, 157, 165  
freedom, xxviii–xxix, xxxi, 47, 59, 60, 62, 172, 267, 296, 305–306, 308, 314, 317–318, 320–321, 330, 349
- Garcia effect, 232  
gene-culture coevolution, 261  
generative grammar, 100, 104, 109, 112n10, 122, 129  
generative model, 207, 215, 216, 219  
genes, 127, 170, 226, 230, 255, 257, 263–267, 269–273, 282, 290–293, 339, 346–347  
genomics, xxx, 255, 268, 269, 282  
Good Tricks, 252, 292, 338, 340  
Guinness (beer), 205–206
- HADD (Hyperactive Agency Detection Device), xxxi, 338  
Hario, v60, 306–307  
Haugeland, John, 14, 257  
heterophenomenology, xx, 98, 167–168, 311  
hippocampus, 68–71, 80–82, 242  
holes, 6, 11, 33, 181  
honeybees, 231, 237, 242, 243  
HOT (Higher-Order Thought), 153–161, 168–169  
Hume, David, 202, 206, 220, 228, 23–237, 251, 343  
Humphrey, Nicholas, 196
- ideology, xxxii, 47, 59, 312, 320n15  
imagery, mental, xxiii, 167, 173–179



- immune system, xxx, 255, 257, 262–267, 290, 292; passive immunity, 263; pathogens, 262–267
- infallibility, 133–134, 139, 143, 144, 146–147, 153–155, 340
- inference, xii–xiii, 41, 84, 134–136, 155–156, 159n18, 160, 309, 338
- information, xv–xix, xxi–xxii, xxv, xxvii, xxx, 7, 9–10, 33n1, 37–38, 44, 49, 52, 59, 63–66, 69–70, 73–74, 77–78, 82, 92, 94, 146, 179, 192, 203, 239, 250, 295–298, 301–303, 307–308, 313–314; information, genetic, 268–273, 292; information, memetic, 274, 278–280, 282, 292–293, 339; information, sensory, 141, 157–158, 174, 177, 180, 181, 183, 187, 188, 197–198, 216, 219, 310
- information processing, 67, 71–72, 80, 81, 202, 203–204, 211
- innovation, xxix, 226, 243, 252, 261, 273, 306
- instrumentalism, 3, 5–6, 12, 17, 27, 341n2, 351–352
- intentionality, original, 18, 46
- interfaces, xviii, 97–99, 128, 316
- interpretivism, xix, 24
- intuition pumps, 95, 139, 169
- inversions, strange, xxxii, 202, 205–207, 213–214, 215
- judgment, xx–xxi, 10n1, 44, 156, 167, 192, 204–205, 207, 213n5, 214, 219, 276, 298, 305–306, 312, 321, 328
- justification, 37, 42, 43–45, 49, 53, 60, 298, 329, 351
- kinds, natural, 51, 52
- Kohler, Wolfgang, 240, 252
- Kosslyn, Steven, 175, 177–179, 307
- Kripke, Saul, 46, 138
- Language of Thought, 3, 58n2, 78, 93
- Law of Effect, xxiv, 225
- learning, instrumental, xxv, xxvii, 231–237, 239, 242, 244–245, 277, 279, 299
- learning, social, 240–241, 243n34
- lexicon, 95, 104–106, 109–112, 114, 115–116, 118–121, 129; lexical rules, 111–114, 122
- magic, xxiv, 8, 19, 63, 127, 172, 187, 189, 207
- manifest image, 6, 16, 47, 172
- McGeer, Victoria, xii, 45, 53
- meaning, 18, 20–26, 29, 33n1, 35, 57–59, 93, 97–99, 106–113, 129, 173, 211, 255, 334, 347, 349; deflationism about, 27–28
- memetics, xxx, 255, 256–262, 267, 269, 273–275, 281–283, 290, 293; memome, xxx, 273
- memory, reconstructive, 65–67, 315–316
- mental time travel, 67–69
- mentalizing, 33, 43, 49, 71, 98n2
- metacognition, xxi, 169
- metaphysics, xxxi, 7, 12, 34–35, 216, 221, 321, 330, 345, 348
- mindset, deliberative, 297–298, 314, 238
- mindset, implementation, 298, 328–329; Implementation intentions, 300–303, 304, 313–314
- mindshaping, xiii, 215, 316
- money, xxx, 11–13, 17, 97
- Monroe, Marilyn, 181–183, 185, 198
- Mr. Clapgras, 213

- Nagel, Thomas, 46, 144, 165
- naturalism, xi, xiv, xxiii, 10, 19, 37, 39–42, 48–49, 53–54, 58n2, 165, 171–172, 175, 179, 189–190, 192–193, 321, 332, 346
- NCC (neural correlate of consciousness), 142n9, 159, 169–170, 190, 192
- neuromarketing, 15
- niche construction, xxvi, xxviii, xxx, 317, 321
- ontology, 3–4, 6–7, 10–11, 13, 17–18, 23–24, 26–29, 33–34, 41, 348
- Ostrom, Elinor, 313n10, 319
- Pacherie, Elisabeth, 297, 301, 304
- Perceptual Symbol System, 278
- personhood, xv–xviii, xx, xxii, xxxi–xxxii, 29, 42, 25, 53, 60, 260, 328, 340–342, 345–346, 351–352
- perspective, first-person, xix, xxiii, xxxii, 15, 133–140, 146–153, 165, 341–342, 351
- phylogenetic tree, xxvi–xxvii, 234
- Pinker, Steven, 114–115, 116, 118, 119, 120, 122
- pluralism, 192, 259n2, 268, 292
- “poker tell,” 168
- postgenomics, 268, 290
- precommitment, 299–300, 304, 309, 314, 319
- prediction, xvii, xxiv, xxv, xxvii, 32, 37, 75, 82, 207–208, 210–212, 219–221, 227, 234–236, 295, 301–302, 320; Predictive Brain Hypothesis, 74n3; 202–216; prediction errors, 208, 210, 212, 214; prediction, quotidian, 40–43, 49–53
- priming, 148–149, 183, 200, 276, 277n15, 300
- pushdown automaton, 251, 297, 307
- qualia, xx–xxiv, 134n2, 138, 150, 158, 166–167, 172, 186, 189–193, 198, 202–205, 206–207, 211–213, 216, 219, 221; Qualia surprise, 214–215, 220
- Quine, W. V. O., 33, 35, 39, 57–58, 93, 136
- Ramachandran, V. S., 181–182, 200
- rationales, free-floating, 34, 59, 304, 310, 321
- rationality, xii–xiv, xv, xvii, 4, 7–9, 14–15, 19, 25, 32, 37–39, 41, 48–50, 52–53, 58, 136, 145, 160, 172, 261, 299, 304, 312n9, 328–329, 338; instrumental rationality, 49–50, 52
- reactivation, sensory, 70–71
- Real Patterns, xv–xvi, 5, 14, 39, 52, 54, 128, 145, 147–148, 166, 216, 340
- reality, literal, 4, 6–7, 17, 20–23, 24, 28
- reality, objective, 4, 147
- regularities, statistical, xviii, xxiv, xxv, 65, 72, 78, 312. *See* Bayesian processes
- religion, 318n13, 331–333, 342–343, 346; religious belief, xxxi–xxxii, 34, 291, 333, 340; religious upbringing, 334–335
- representation, mental, xviii–xix, xxi, xxii–xxiii, xxvi, xxvii, xxx, 44, 65, 78–82, 92, 95–123, 127–128, 129, 174–175, 177–178, 180–181, 183, 185–186, 188, 191, 193, 196–199, 235n18, 242, 274, 278–280, 293, 299, 301, 303, 307, 315–316, 350; representation, external, 308, 310–312
- retrieval, probabilistic, xviii, 65, 72, 74, 77–78, 83
- reverse engineering, xviii, xxxi, 33, 63, 67, 78, 83, 98

- reward, xxiv–xxvi, xxix, xxx, 228, 231, 232–234, 239, 240n29, 244, 251, 277n15, 295, 299, 304–305, 311, 312, 352
- Ryle, Gilbert, 39, 49, 83–84, 338
- Sam (the art critic), 46, 57
- scaffolding, social and material, 307–309
- schemas, syntactic, 101, 108
- schemata, 64–65, 74–77, 101–103, 108, 112–114, 116–118, 120–122, 129, 215, 274
- self-interpretation, 144–145
- Sellars, Wilfrid, 16, 19, 29, 136
- simulation, 68–69, 72, 83, 278–280
- skyhooks, 250, 290
- spike trains, 79, 203, 205
- Spinoza, Benedict de, 300, 304, 306, 321
- squirrel, albino, 214–215, 220
- stances. ontology of, 3–16; The boxing stance, 8; The clinical stance, 9, 19; The Design stance, 3, 7, 8, 22; The economic stance, 13; The intentional stance, xii–xiv, xv–xviii, xxxii, 3, 4–7, 8, 15, 17, 23, 32–34, 36–43, 49–52, 62, 66, 78, 82, 136–137, 144–148, 150, 151, 153, 156, 159–161, 165–166, 167, 255, 328, 337–340, 349–351; The interpretive stance, 24–28; The personal stance, 49n3, 53, 60, 341; The physical stance, 3, 16, 22, 136, 339–340, 349–350; The teleological stance, 49–50
- statistics, ensemble, 198–199
- subpersonal cognitive psychology, xi, xv–xix, xx–xxii, 15, 39, 66–67, 78–79, 82, 92–93, 128, 186, 209, 211, 212, 213, 295n1, 310, 314, 317, 330
- Tollefsen, Deborah, 304, 310–311, 312, 315
- Tomasello, Michael, 108, 120, 238, 241
- Tower of Generate-and-Test, xxv–xxvii, 226, 250, 303; Carnapian creatures, xxvii, 237–240, 244, 252; Darwinian creatures, xxv, xxvi–xxvii, 225, 226–227, 233, 243, 250; Gregorian creatures, xxvi, xxvii, xxix, 226, 227, 240, 250, 252–253; Humean creatures, xxvii, 227–230, 232–233, 238, 243, 244, 251; Leibnizian creatures, 230; Mimetic creatures, xxvi, 241; Pearlian creatures, xxvii, 238–239, 240, 252; Popperian creatures, xxvi–xxvii, 226, 237, 239–240, 244, 250, 252; Skinnerian creatures, xxv, xvii, 226–227, 231–234, 237, 238, 239, 240, 241, 243–244, 250–251; Tolmanian creatures, xxvii, 242, 252; Vygotskian creatures, 242–243
- transactive processing, 314–316
- transformism, 256
- “tuned deck,” 192
- Turing Machines, xvii, 253
- turnstiles, 186–188
- unification, 102, 109, 122
- vagueness, ontological, 337
- value, xii–xiii, xxix, 46, 261, 295, 296, 299, 301, 303, 304–306, 312, 320–321, 329, 334, 341–342, 347, 349

- variables, latent, 211, 215, 216, 219
- variation, 50, 75, 102, 103, 122, 225, 239, 264–267, 271, 292
- vectorwaves, 190–192, 199
- Wittgenstein, Ludwig, 20, 161, 173
- working memory, xxi, xxii, 158, 183, 187–188, 191, 200, 308
- worlds, notional, xxvi, xxix, 311–312, 317

