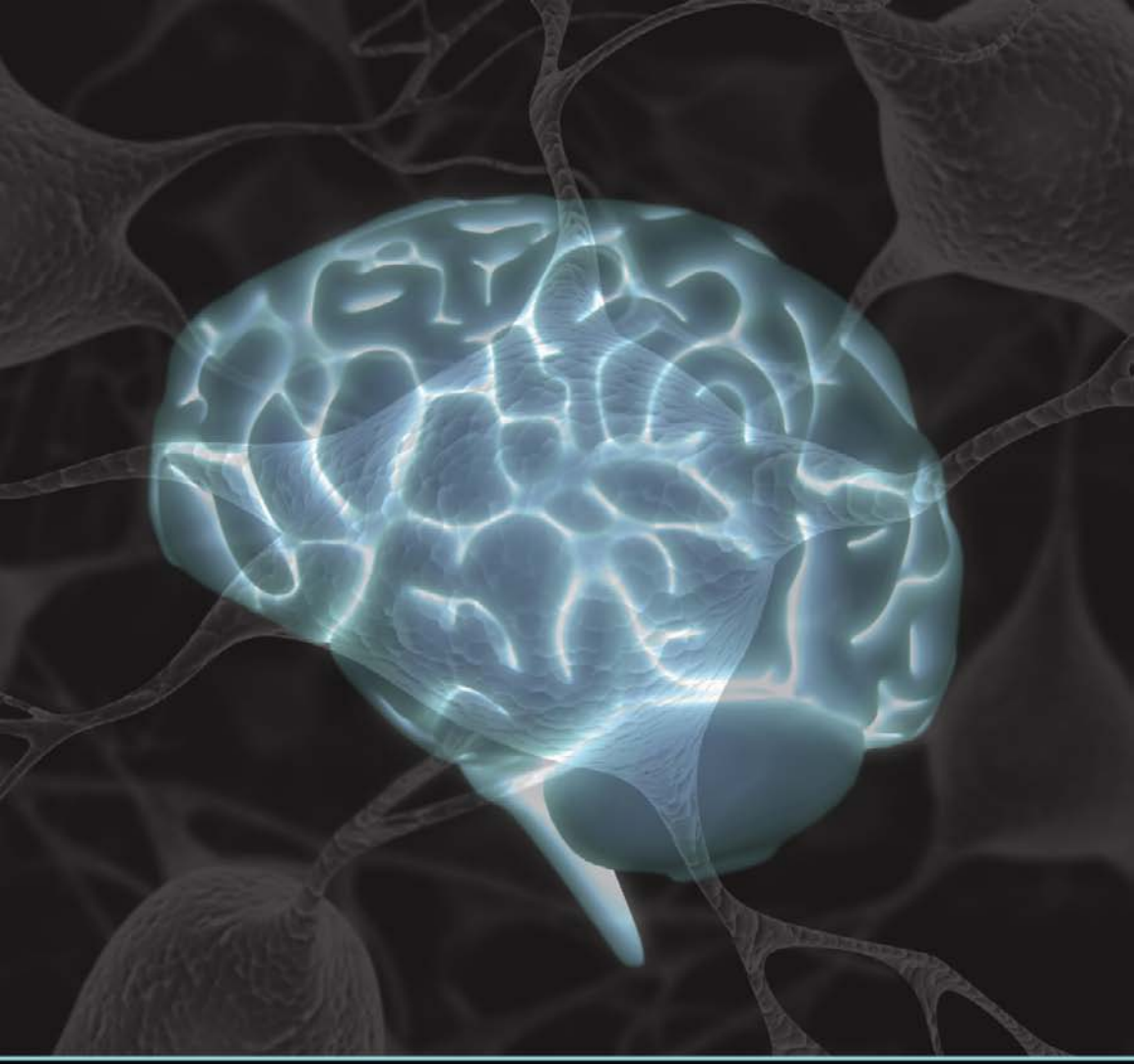


The Consciousness Paradox

Consciousness, Concepts, and Higher-Order Thoughts

Rocco J. Gennaro

The Consciousness Paradox

Representation and Mind

Hilary Putnam and Ned Block, Editors

Representation and Reality, Hilary Putnam
Explaining Behavior: Reasons in a World of Causes, Fred Dretske
The Metaphysics of Meaning, Jerrold J. Katz
A Theory of Content and Other Essays, Jerry A. Fodor
The Realistic Spirit: Wittgenstein, Philosophy, and the Mind, Cora Diamond
The Unity of the Self, Stephen L. White
The Imagery Debate, Michael Tye
A Study of Concepts, Christopher Peacocke
The Rediscovery of the Mind, John R. Searle
Past, Space, and Self, John Campbell
Mental Reality, Galen Strawson
Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind, Michael Tye
Representations, Targets, and Attitudes, Robert Cummins
Starmaking: Realism, Anti-Realism, and Irrealism, Peter J. McCormick
A Logical Journey: From Gödel to Philosophy, Hao Wang
Brainchildren: Essays on Designing Minds, Daniel C. Dennett
Realistic Rationalism, Jerrold J. Katz
The Paradox of Self-Consciousness, Jose Luis Bermudez
In Critical Condition: Polemical Essays on Cognitive Science and the Philosophy of Mind, Jerry Fodor
Mind in a Physical World: An Essay on the Mind–Body Problem and Mental Causation, Jaegwon Kim
The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology, Jerry Fodor
New Essays on Semantic Externalism and Self-Knowledge, Susana Nuccetelli
Consciousness and Persons: Unity and Identity, Michael Tye
Naturalistic Realism and the Antirealist Challenge, Drew Khlentzos
Wittgenstein and the Moral Life: Essays in Honor of Cora Diamond, Alice Crary, editor
Time and Realism: Metaphysical and Antimetaphysical Perspectives, Yuval Dolev
Austere Realism: Contextual Semantics Meets Minimal Ontology, Terence E. Horgan and Matjaž Potrč
Consciousness Revisited: Materialism without Phenomenal Concepts, Michael Tye
The Consciousness Paradox: Consciousness, Concepts, and Higher-Order Thoughts, Rocco J. Gennaro

The Consciousness Paradox

Consciousness, Concepts, and Higher-Order Thoughts

Rocco J. Gennaro

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

For information about special quantity discounts, please e-mail special_sales@mitpress.mit.edu

This book was set in Stone Sans and Stone Serif by Graphic Composition, Inc.

Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Gennaro, Rocco J.
The consciousness paradox : consciousness, concepts, and higher-order thoughts / Rocco J. Gennaro.
p. cm.—(Representation and mind series)
"A Bradford Book."
Includes bibliographical references (p.) and index.
ISBN 978-0-262-01660-5 (hardcover : alk. paper)
1. Consciousness. 2. Concepts. 3. Thought and thinking. I. Title.
B808.9.G46 2012
126—dc22

2011011586

10 9 8 7 6 5 4 3 2 1

To Deidra, Olivia, and Joseph

Contents

Acknowledgments ix

1 Introduction 1
- 1.1 The Problem 1
- 1.2 The Plan 3
- 1.3 Some Terminology and Distinctions 5

2 In Defense of the HOT Thesis 11
- 2.1 Varieties of Representationalism 11
- 2.2 Defending Reductive Representationalism 15
- 2.3 Consciousness and Intentionality 21
- 2.4 HOT Theory: An Initial Defense 28
- 2.5 More on Mental Content 33

3 Assessing Three Close Rivals 39
- 3.1 First-Order Representationalism (FOR) 39
- 3.2 Dual-Content Theory 45
- 3.3 Higher-Order Perception (HOP) Theory 49

4 From HOT Theory to the Wide Intrinsicality View 55
- 4.1 A False Dilemma 55
- 4.2 Misrepresentation: A First Pass 59
- 4.3 The Problem of the Rock 70
- 4.4 The Hard Problem of Consciousness 75
- 4.5 Objections and Replies 88

5 Against Self-Representationalism 103
- 5.1 Three Views of State Consciousness 104
- 5.2 Against Pure Self-Referentialism 104
- 5.3 Another Approach: Peripheral Awareness 116
- 5.4 Three More Attempts and a Counterargument 126

6 In Defense of Conceptualism 135
6.1 What Is Conceptualism? 135
6.2 HOT Theory and Conceptualism 147
6.3 The Richness of Conscious Experience 161
6.4 Fineness of Grain 173

7 Concept Acquisition and Infant Consciousness 185
7.1 The Real Hard Problem 185
7.2 Innateness 186
7.3 Concept Acquisition 199
7.4 Conceptualism and Concept Acquisition 210
7.5 HOT Theory and Infant Consciousness 219

8 Animal Consciousness 229
8.1 Carruthers, Animals, and HOT Theory 229
8.2 Animals and I-Thoughts 237
8.3 Lloyd Morgan's Canon and Parsimony 252
8.4 An Aside on Autism 257
8.5 Animals and Conceptualism: The Continuity Argument 262

9 Into the Brain 269
9.1 The Neural Correlates of Consciousness (NCCs) 269
9.2 How Global Is HOT Theory? 276
9.3 Parts, Wholes, and Feedback Loops 282
9.4 The Binding Problem and the Unity of Consciousness 291
9.5 Conclusion of the Book 302

Notes 305
References 329
Index 371

Acknowledgments

Most of this book is new material. Chapters 2, 3, 6, 7, and 9 are almost entirely new. Some of chapter 8 is a much expanded version of Gennaro 2009, noted below. Sections 8.4 and 8.5 are entirely new. Larger portions of chapters 4 and 5 have been previously published. However, in each case, there is significant expansion, updating, reworking, and reorganization. The previous publications in question are the following:

"Animals, Consciousness, and I-Thoughts," in *Philosophy of Animal Minds*, ed. Robert Lurz (New York: Cambridge University Press, 2009), 184–200.
"Representationalism, Peripheral Awareness, and the Transparency of Experience," *Philosophical Studies* 139 (May 2008): 39–56.
"Review of Peter Carruthers' *Consciousness: Essays from a Higher-Order Perspective*," *Psyche* 12 (August 2006), http://theassc.org/files/assc/2645.pdf.
"Between Pure Self-Referentialism and the (Extrinsic) HOT Theory of Consciousness," in *Self-Representational Approaches to Consciousness*, ed. Uriah Kriegel and Ken Williford (Cambridge, MA: MIT Press, 2006), 221–248.
"The HOT Theory of Consciousness: Between a Rock and a Hard Place?" *Journal of Consciousness Studies* 12, no. 2 (February 2005): 3–21.
"Higher-Order Thoughts, Animal Consciousness, and Misrepresentation: A Reply to Carruthers and Levine," in *Higher-Order Theories of Consciousness*, ed. Rocco J. Gennaro (Amsterdam: John Benjamins, 2004), 45–66.
"Higher-Order Theories of Consciousness: An Overview," introduction to *Higher-Order Theories of Consciousness*, ed. Rocco J. Gennaro (Amsterdam: John Benjamins, 2004), 1–13.

I thank Cambridge University Press, the MIT Press, John Benjamins Publishers, *Journal of Consciousness Studies*, *Philosophical Studies*, and *Psyche* for reprint permission.

Over the years, I have benefited greatly from discussions, in person and via e-mail, with many others, including Colin Allen, John Beeckmans,

José Bermúdez, John Bickle, Ned Block, Andy Brook, Richard Brown, Alex Byrne, Peter Carruthers, David Chalmers, Fred Dretske, Don Dulany, Rob Goldstone, Valerie Hardcastle, Chris Hill, Douglas Herrmann, Terry Horgan, Uriah Kriegel, Joe Levine, Bill Lycan, Robert Lurz, Pete Mandik, Jesse Prinz, David Rakison, Bill Robinson, David Rosenthal, Laurie Santos, Bill Seager, Elizabeth Schechter, Charles Siewart, Robert Van Gulick, Josh Weisberg, and Ken Williford. I owe the most in recent years to David Rosenthal, Uriah Kriegel, and Bill Lycan for their thought-provoking work and for numerous helpful discussions and e-mail exchanges. Special thanks also to Bill Robinson for particularly detailed comments on chapters 6 and 7. I also wish to thank four anonymous referees for the MIT Press, whose comments and suggestions greatly improved the final product.

I am thankful to the University of Southern Indiana for a stimulating Summer Writing Institute in May 2010 and to Indiana State University for a sabbatical leave before my move to Southern Indiana in July 2009.

Finally, I am grateful to Phil Laughlin at the MIT Press for taking up this project and helping to bring it to completion. Thanks also to Katie Persons and Judy Feldmann for their editorial work.

1 Introduction

1.1 The Problem

Consciousness is arguably the most important area within contemporary philosophy of mind, with an explosion of research over the past thirty years from philosophers, psychologists, and scientists.[1] Consciousness is also perhaps the most puzzling aspect of the world, and yet it is so very familiar to each of us. Attempts to explain it in neurophysiological or even cognitive terms are still met with great resistance. It seems to many that conscious mental states simply cannot be reduced to, or explained in terms of, something less problematic. In this book, I defend and further develop a metapsychological reductive representational theory of consciousness and then apply it to several importantly related problems, including concept acquisition and animal consciousness.

Going back to my book *Consciousness and Self-Consciousness* (1996), I have defended a version of the higher-order thought (HOT) theory of consciousness, which says that what makes a mental state conscious is that there is a suitable higher-order thought directed at the mental state. Higher-order thoughts (HOTs) are metapsychological or metacognitive states, that is, mental states directed at other mental states. HOT theory is primarily concerned with explaining how conscious mental states differ from unconscious mental states. It seems reasonable to think that conscious mental states are states that we are "aware of" in some sense. Its best-known defender is David Rosenthal.[2]

I called my version of HOT theory the wide intrinsicality view (WIV) for reasons we will see in due course. Moreover, in Gennaro 1996, I was chiefly concerned to argue for the more general Kantian thesis that consciousness entails self-consciousness. Defending HOT theory was therefore mainly a means to that end. Since that book, however, I have further developed my own version of HOT theory (Gennaro 2006a), including attention to issues

such as animal consciousness (Gennaro 2004a, 2009) and the well-known "hard problem" of consciousness (Gennaro 2005b). In some cases, I have simply defended HOT theory against a specific objection (Gennaro 2003).[3]

In addition to further defending HOT theory, however, I am interested in solving what I take to be a larger underlying paradox that I will call the Consciousness Paradox, namely, how it is possible to hold the following set of apparently inconsistent yet independently plausible theses:

1. *The HOT Thesis*: A version of the HOT theory is true (and thus a version of reductive representationalism is true).
2. *The Hard Thesis*: The hard problem of consciousness, that is, the problem of explaining exactly how or why subjective experiences are produced from brain activity (or from any combination of unconscious mental activity), can be solved.
3. *The Conceptualism Thesis*: Conceptualism is true, that is, all conscious experience is structured by concepts possessed by the subject.
4. *The Acquisition Thesis*: The vast majority of concepts are acquired, though there is a core group of innate concepts.
5. *The Infants Thesis*: Infants have conscious mental states.
6. *The Animals Thesis*: Most animals have conscious mental states.
7. *The HOT-Brain Thesis*: There is a plausible account of how HOT theory, and especially the WIV, might be realized in the brain and can lead to an informative neurophysiological research agenda. Alternatively: HOT theory is interestingly related to and consistent with a number of leading empirical theories of consciousness.

Indeed, it is often claimed that HOT theory alone is inconsistent with several of the other theses in the foregoing list. For example, some have argued that the HOT Thesis conflicts with the Animals Thesis because animals cannot have what seem to be fairly sophisticated HOTs of the form "I am in mental state M now." Much the same has been said about the Infants Thesis. Further, the Conceptualism Thesis has been thought by some to contradict the Acquisition and Animals Theses. Do animals even possess concepts? If all conscious perceptual experience is determined by already possessed concepts, then how could we acquire new concepts? Although I think that each of the theses is independently plausible and defensible, some are more controversial than others, and it is necessary to explain how they can all be mutually consistent. Thus my overall aim is to argue for a philosophical theory of consciousness while applying it to other significant issues such as concept possession and concept acquisition, topics not frequently found in the philosophical literature on consciousness. This book

also addresses interdisciplinary topics such as animal consciousness and how HOT theory might be realized in the brain. Thus I hope that it will be of interest to nonphilosophers as well as philosophers.

Most cognitive science and empirical works on, for example, concepts or animal consciousness do not address central philosophical theories of consciousness. Some are primarily experimental or scientific works by authors not necessarily interested in consciousness research as such.[4] These works mostly focus on the nature of concepts, concept acquisition, and theories of mental state attribution, but without delving much into the philosophical problem of consciousness. On the other hand, many of the more philosophical works do not integrate a specific theory of consciousness with the cognitive science literature on the topics listed earlier.

There are of course numerous important and helpful anthologies in the field. Some are specifically on HOT theory or closely related theories of consciousness (Gennaro 2004b; Carruthers 2005; Rosenthal 2005; and Kriegel and Williford 2006), with the volumes by Rosenthal and Carruthers representing collections of their own writings. Finally, although there are many other excellent anthologies on consciousness, they are obviously not designed to put forth a single unified theory.[5] The present work therefore addresses various problems in novel ways and in relation to HOT theory.

1.2 The Plan

It might be useful to think of this book as comprising two parts, chapters 1 through 5 and chapters 6 through 9. In the present chapter, I lay out the overall problem and make some key distinctions. In chapters 2 through 5, I defend the HOT Thesis. I also argue for the Hard Thesis in chapter 4. In chapters 6 through 9, I defend the remaining theses paying special attention to how each thesis is consistent with the others, including the HOT Thesis.

In chapter 2, I first defend representationalism, which is the thesis that phenomenal properties are identical to certain representational properties. I then argue that *reductionist* representationalism is most desirable if we are to explain consciousness. I defend the associated view that intentionality is more primitive than consciousness, and offer an account of mental content. Finally, I present an initial defense of HOT theory.

In chapter 3, I argue against several theories of consciousness that are somewhat close relatives to HOT theory. I reject first-order representationalism (FOR) such as the account offered by Michael Tye (1995, 2000). I argue that FOR cannot adequately explain the difference between conscious and

unconscious mental states. I also argue that HOT theory is preferable to the higher-order perception (HOP) alternative proposed by William Lycan (1996) and the dual-content (or dispositional HOT) theory of Peter Carruthers (2000, 2005).

In chapter 4, I first motivate the need for a modified version of Rosenthal's HOT theory, namely, the wide intrinsicality view (WIV), focusing on two major objections to HOT theory: the so-called problem of the rock and the misrepresentation objection. I then take on the hard problem of consciousness and end with replies to several further objections to the WIV. I argue that HOT theory is immune to Chalmers's (1995) criticisms of other attempted reductionist accounts of consciousness, but also that a similar version of the problem can be solved.

In chapter 5, I turn my attention to another related theory of consciousness. Although there is something right about the idea that conscious mental states represent *themselves*, I argue against the so-called self-representational theory of consciousness. Prominent in this chapter is critical discussion of recent work by Uriah Kriegel.[6]

At this point, I conclude that the HOT and Hard Theses have been tentatively established though much more will follow in later chapters especially with regard to how the HOT Thesis is consistent with the remaining theses.

Chapter 6 is devoted to defending the Conceptualism Thesis against some well-known objections and several prominent critics, such as Sean Kelly and Christopher Peacocke. The central issue is whether or not one can have conscious experience of objects or properties without having the corresponding concepts. In this chapter, I offer a more detailed account of concept possession and also explain crucial connections between HOT theory and conceptualism.

In chapter 7, I first defend the Acquisition Thesis. One objection to both HOT theory and conceptualism is that concept acquisition is impossible or extremely difficult to explain. For example, how can one acquire the concept of a novel (type of) object if having conscious perception of that object already presupposes that the subject possesses the concept? This chapter addresses this problem, which I take to be the "real hard problem" of consciousness. With the aid of a wealth of experimental evidence from the infant and child developmental literature, I argue that concept acquisition is indeed consistent with both conceptualism and HOT theory. I also defend the Infants Thesis and argue that it is consistent with the HOT and Acquisition Theses.

In chapter 8, I defend the Animals Thesis. The nature of concept possession and so-called "I-thoughts" plays an important role in the ever-

increasing animal cognition literature. In previously defending HOT theory, I have responded at length to the allegation that HOTs (along with their constituent concepts) are too sophisticated for many animals to have (Gennaro 1993, 2004a, 2009). I continue this defense in chapter 8 and then extend it to include discussion of conceptualism. It is also necessary to examine experimental results to determine whether or not animals have self-concepts, self-awareness, episodic memory, and an ability to attribute mental states to others. I argue that the Animals Thesis is consistent with both the Conceptualism and HOT Theses.

Chapter 9 contains a defense of the HOT-Brain Thesis. This chapter explores the neurophysiological evidence for HOT theory and, especially, the wide intrinsicality view (WIV). I argue that my theory can shed light on, and is consistent with, several empirical theories of consciousness. I examine the literature on the "neural correlates of consciousness" (NCCs), that is, the ongoing scientific project of determining the precise neural correlates of consciousness. Additional motivation for the HOT-Brain Thesis is, for example, to refute Antti Revonsuo's recent charge that "these theories [i.e., HOT theories] have not had any major impact on the empirical study of consciousness" (2010, 189). The well-known "binding problem" must also be addressed in this context; namely, how does the varied incoming information to the brain result in a unified and coherent visual experience? The chapter also critically discusses the closely linked problem of the unity of consciousness.

Overall, I claim that the Consciousness Paradox can be solved.

1.3 Some Terminology and Distinctions

The concept of consciousness is notoriously ambiguous. It is important first to make several distinctions and to define key terms. The abstract noun "consciousness" is not often used in the contemporary literature, though it originally derives from the Latin *con* (with) and *scire* (to know). One can have knowledge of the external world or one's own mental states through consciousness. The primary contemporary interest lies more in the use of the expressions "x is conscious" or "x is conscious of y." Under the former category, perhaps most important is the distinction between *state* and *creature* consciousness (Rosenthal 1993a). We sometimes speak of an individual mental state, such as a pain or perception, as being conscious. On the other hand, we also often talk about organisms or creatures as conscious, such as when we say that "human beings are conscious" or "cats are conscious." Creature consciousness is simply meant to refer to the fact that an organism

is awake, as opposed to sleeping or in a coma. However, some kind of state consciousness is normally implied by creature consciousness; that is, if a creature is conscious, then it must have conscious mental states.

Due to the lack of a direct object in the expression "x is conscious," this is usually referred to as *intransitive* consciousness, in contrast to *transitive* consciousness, where the locution "x is conscious of y" is used (Rosenthal 1993a). Most contemporary theories of consciousness are aimed at explaining *state consciousness*, that is, what makes a mental state conscious. This is also the case for HOT theory in the sense that intransitive (state) consciousness is explained in terms of transitive consciousness.

It might seem that the term "conscious" is synonymous with, say, "awareness" or "experience" or "attention." However, it is crucial to recognize that this is not generally accepted today. For example, though perhaps somewhat atypical, one might hold that there are unconscious *experiences*, depending, of course, on how the term "experience" is defined (Carruthers 2000). More common is the belief that we can be *aware* of external objects in some unconscious sense, such as during instances of subliminal perception. The expression "conscious awareness" does not therefore seem to be redundant. Finally, it is not clear that consciousness ought to be restricted to *attention*. It seems plausible to suppose, for example, that one is conscious of objects to some extent in one's peripheral visual field even though one is attending to a more narrow (or focal) set of objects within that visual field. Needless to say, contemporary philosophers and psychologists are nearly unanimous in allowing for unconscious *mental states* or *representations*, though they differ as to whether this applies to *all* kinds of mental states including, say, pains and feelings.

Perhaps the most fundamental and commonly used notion of "conscious" is captured by Thomas Nagel's famous "what it is like" sense (Nagel 1974). When I am in a conscious mental state, there is "something it is like" *for me* to be in *that state* from the subjective or first-person point of view. When I smell a rose or have a conscious visual experience, there is something it "seems" or "feels like" from my perspective. An organism such as a bat is conscious if it is able to experience the world through its echolocation senses. There is also something it is like to be a conscious creature, whereas there is nothing it is like to be a table or tree. This is primarily the sense of "conscious state" that I use throughout the book. "What it's like" basically means "how a conscious state is for the subject." When it comes to capturing the main *phenomenon to be explained*, it seems to me that we most often have Nagel's "something it is like" sense in mind.

There are still, though, a cluster of expressions and technical terms associated with Nagel's sense, and some authors simply stipulate the way that they use them. For example, philosophers often refer to conscious states as *phenomenal* or *qualitative* states. More technically, philosophers frequently describe such states as having qualitative properties called "qualia" (singular, *quale*). Chalmers explains that a "mental state is conscious if there is something it is like to be in that mental state. . . . We can say that a mental state is conscious if it has a qualitative feel. . . . These qualitative feels are also known as phenomenal qualities, or qualia for short" (1996, 4). There is significant disagreement over the nature, and even the existence, of qualia, but they are most often understood as the felt properties or qualities of conscious states. There is something it is like to have qualia or to be in a qualitative state. Most generally, perhaps, qualia are "introspectively accessible, phenomenal aspects of our mental lives" (Tye 2009a). But even this can be misleading if it is taken to imply that only introspected, or introspectible, states have qualia. Surely first-order, or world-directed, conscious states also have qualia. Kind (2008) explains that "qualia are subjective or qualitative properties of experiences. What it feels like, experientially, to see a red rose is different from what it feels like to see a yellow rose. Likewise for hearing a musical note played by a piano and hearing the same musical note played by a tuba. The qualia of these experiences are what give each of them its characteristic 'feel' and also what distinguish them from one another." In any case, qualia are most often treated as properties of some mental states, though some use the term "qualia" in the more external sense of "the qualities of *what is represented*." I will use it in the former sense.[7]

One also finds closely allied expressions like "phenomenal character" and "subjective character" in the literature. Tye (2009a), for example, tells us that the "phenomenal character of an experience is what it is like subjectively to undergo the experience." More explicitly, Kriegel (2009a) is at great pains to distinguish what he calls "qualitative character" from "subjective character" under the larger umbrella of "phenomenal character" because they play such a central role in his theory of consciousness. He explains that "a phenomenally conscious state's qualitative character is what makes it the phenomenally conscious state it is, while its subjective character is what makes it a phenomenally conscious state at all" (Kriegel 2009a, 1). In his view, then, the *phenomenally conscious* experience of the blue sky should be divided into two components: (1) its *qualitative character*, which is the "bluish" component of the experience (or the *what* of the experience), and (2) its *subjective character*, which is what he calls the "for-me" component (or what determines *that* it is conscious). As we will see in chapter 5, I think

that Kriegel is mistaken in thinking that subjective character is itself phenomenally conscious, though I am more sympathetic with his account of qualitative character.[8]

Finally, Ned Block (1995) makes an oft-cited distinction between *phenomenal* consciousness (or "phenomenality") and *access* consciousness. Phenomenal consciousness is very much in line with Nagel's notion described earlier. However, Block defines the quite different notion of access consciousness in terms of a mental state's relationship with other mental states, for example, a mental state's "availability for use in reasoning and rationality guiding speech and action" (Block 1995, 227). This view would, for example, count a visual perception as (access) conscious not because it has the "what it's likeness" of phenomenal states but because it carries visual information that is generally available for use by the organism, regardless of whether or not it has any qualitative properties. Access consciousness is therefore a functional notion concerned with what such states do. Although something like this idea is certainly important in cognitive science and philosophy of mind generally, not everyone agrees that access consciousness deserves to be called "consciousness" in any important sense. Block himself argues that neither sense of consciousness implies the other, while others urge that a more intimate connection holds between the two. For example, according to HOT theory, phenomenality would entail (higher-order) access consciousness, but not all access consciousness would have phenomenality.

My sense is that some authors only add to the terminological confusion by introducing new (or not so new) distinctions into the literature instead of clarifying existing meanings of "consciousness" or simply adopting a prior definition over others. Has, for example, Block's distinction between access and phenomenal consciousness really *clarified* the matter? It is important to resist the constant temptation to introduce our own special terminology, though we are all perhaps guilty to some extent. For example, here is a sample from my own 1996 book: "Qualia are the properties of phenomenal states that determine their qualitative character, i.e. 'what it is like' to have them" (Gennaro 1996, 7). A *phenomenal state* can occur unconsciously and is "a mental state which . . . typically has qualitative properties" (7), whereas a *qualitative state* is a "phenomenal state with a qualitative property" (8). Thus, according to my 1996 view, a phenomenal state can be unconscious, but a qualitative state must be conscious.[9]

I make a plea for more uniform usage whenever possible. Unless I specifically indicate otherwise, I will from now on use the terms "phenomenal," "qualitative," and "experience" as conscious in Nagel's sense, but I will

allow for unconscious *awareness* and unconscious *representations* directed at the outer world or one's own mental states. So, for me, there are no unconscious experiences and, in contrast to other higher-order theorists, no unconscious qualitative states (or unconscious qualia). There is little reason to have an unconscious counterpart for *each* of the terms I have listed. I will also avoid using as much of the foregoing technical jargon as possible throughout this book where I can do so without sacrificing rigor or accuracy. However, it will sometimes be necessary to get into the terminological weeds, especially when discussing other views. In any case, I now turn to a defense of the HOT thesis.

2 In Defense of the HOT Thesis

In this chapter, I begin a defense of the HOT Thesis, namely, that a version of the HOT theory is true and thus a version of reductive representationalism is true. This first involves explaining several flavors of representationalism (sec. 2.1), as well as making a case for a reductionist approach to consciousness (sec. 2.2). In section 2.3, I argue that intentionality is prior to consciousness partly via a critical examination of Searle's well-known Connection Principle. In section 2.4, I offer an initial defense of HOT theory. Finally, in section 2.5, I explore further the nature of mental content in light of HOT theory.

2.1 Varieties of Representationalism

Some current theories of consciousness attempt to reduce it to mental representations of some kind. The notion of a representation is, of course, extremely general and can be applied to photographs, signs, and various natural objects, such as the rings inside a tree. Much of what goes on in the brain, however, might also be understood in a representational way, for example, as mental events representing outer objects partly because they are caused by those objects. Philosophers often call these states *intentional states* that have representational content, that is, mental states that are "about" or "directed at" something such as a thought about a house or a perception of a tree.

The view that we can explain conscious mental states in terms of representational or intentional states is called *representationalism* (or intentionalism). Although not automatically reductionist in spirit, most versions of representationalism do indeed attempt such a reduction. Most representationalists, such as higher-order (HO) theorists, think that there is then room for a "second-step" reduction to be filled in later by neuroscience. One motivation for representationalism is that a naturalistic account of intentionality

can arguably be more easily attained, such as via causal theories whereby mental states are understood as representing outer objects by virtue of some reliable causal connection. The idea, then, is that if consciousness can be explained in representational terms and representation can be understood in purely physical terms, then there is the promise of a naturalistic theory of consciousness. A representationalist will typically hold that the qualitative properties of experience, or qualia, can be explained in terms of the experiences' representational properties. The claim is that conscious mental states have no mental properties other than their representational properties. Two conscious states with all the same representational properties will not differ phenomenally. For example, when I look at the blue sky, what it is like for me to have a conscious experience of the sky is simply identical with my experience's representation of the blue sky.

I cannot fully survey here the dizzying array of representationalist positions (Chalmers 2004; Lycan 2005). I believe that the most plausible form of representationalism is what has been called *strong representationalism*. It is basically the view that having representations of a certain kind suffices for having qualia and thus for conscious mental states. It is sometimes contrasted with *weak* representationalism, which is the view that conscious experience always has representational content of some kind.

It is also important at the outset to distinguish the *content* of a mental state from the *state* or *vehicle* that has the content. This is the difference between what is represented, or what the state is about, and what is doing the representing. Two other pairs of distinctions involve how best to characterize, first, the mental contents in question and, second, the kinds of properties represented.

(1) *Wide* representationalism holds that "both phenomenal properties and the representational properties they are equivalent to are taken to depend on a subject's environment" (Chalmers 2004, 165). This is the view of most representationalists, including Dretske (1995), Tye (1995), and Lycan (1996). It has its roots in the literature on propositional attitudes, such as beliefs and thoughts, which has been taken to show that two physically identical subjects with different environments will have different mental contents (Putnam 1975). For example, a belief about water on Earth will be about H_2O, whereas it will be about XYZ on "Twin Earth." The main idea is that the content (or meaning) of one's mental states depends on one's environment. In contrast, *narrow* representationalism is the view that phenomenal properties, and the representational properties they are equivalent to, depend on a subject's internal state, so that molecular duplicates will necessarily share mental contents. Narrow representationalists think that

molecular duplicates share something significant, even if there are other differences when the relevant mental states are individuated widely.

(2) Within narrow and wide representationalism, one might also disagree about what kinds of properties are represented. For example, one natural way to think of mental content involves objects and properties in the world. Following Russell, they have been called *Russellian contents* (Chalmers 2004). The concepts involved in a belief, for example, have *extensions*, namely, objects and properties that are picked out by the concepts. If I believe that Venus is the second-closest planet to the Sun, then my belief is directed at Venus. On the other hand, following Frege, one might suppose that there are also *Fregean contents*. Mental contents are composed of concepts, which not only have extensions but also have *modes of presentation*, or what might best be described as "a way of thinking about the referent." This mirrors Frege's well-known distinction between reference and sense. So, according to this view, the belief about Venus also has the mode of presentation "second planet to the Sun," which is *the way* that I am conceiving of Venus in that case. Fregean content can differ while the Russellian content remains fixed. I may alternatively believe that Venus is the Morning Star, which involves a different mode of presentation. I return to these distinctions in section 2.5.

For now, it is worth briefly introducing three common flavors of representationalism, each of which I discuss at greater length in later chapters. The central question that should be answered by any theory of consciousness is: What makes a mental state a conscious mental state? That is, what differentiates unconscious mental states from conscious mental states?

2.1.1 First-Order Representationalism (FOR)

First-order representational theories of consciousness refer to theories that attempt to explain conscious experience in terms of world-directed (or first-order) intentional states. Two frequently cited FO theories are those of Dretske (1995) and Tye (1995, 2000), though there are many others as well (Byrne 2001; Thau 2002; Droege 2003). Like other FO theorists, Tye holds that the representational content of my conscious experience (that is, what my experience is directed at) is identical with the phenomenal properties of experience. Aside from reductionistic motivations, Tye and others often invoke the notion of the *transparency of experience* to support their view (Harman 1990). This argument derives from Moore (1903) and is based on the phenomenological first-person observation that when one turns one's attention away from, say, the blue sky and onto one's experience itself, one is still only aware of the blueness of the sky. The experience itself is

not blue; rather, one "sees right through" the experience to its representational properties, and thus there is nothing else to one's experience over and above such properties.

As we will see in chapter 3, FO theorists believe that much the same goes for all kinds of conscious states, including pains and emotions.

2.1.2 Higher-Order Representationalism (HOR)

Another tradition has attempted to understand consciousness in terms of higher-order awareness. For example, some cite John Locke (1689/1975), who once said that "consciousness is the perception of what passes in a man's own mind." This is a bit misleading because, unlike HO theorists, Locke did not believe in unconscious thoughts.[1] In general, the idea is that what makes a mental state conscious is that it is the object of some kind of higher-order representation (HOR). A mental state M becomes conscious when there is a HOR of M. A HOR is a metapsychological or metacognitive state, that is, a mental state directed at another mental state. So, for example, my desire to write a good book becomes conscious when I am (noninferentially) "aware of" the desire. Intuitively, it seems that conscious states, as opposed to unconscious ones, are mental states that I am aware of in some sense. Any theory that attempts to explain consciousness in terms of higher-order states is known as a higher-order (HO) theory of consciousness. HO theories thus attempt to explain consciousness in mentalistic terms, that is, by reference to notions such as "thoughts" and "awareness." We might say that conscious mental states arise when two *un*conscious mental states are related in a certain specific way, namely, when one of them (the HOR) is directed at the other (M).

There are various kinds of HO theory, with the most common division between higher-order *thought* (HOT) theories and higher-order *perception* (HOP) theories. HOT theorists, such as David Rosenthal (1997, 2005), think it is better to understand the HOR as a thought of some kind. HOTs are treated as *cognitive* states involving conceptual components. HOP theorists urge that the HOR is instead a *perceptual* or *experiential* state (Lycan 1996) that does not require the kind of conceptual content invoked by HOT theorists. Although HOT and HOP theorists agree on the need for a HO theory of consciousness, they also often argue for the superiority of their respective positions (Lycan 2004; Rosenthal 2004).

2.1.3 Hybrid Representational Views

A related group of representational theories holds that the HOR in question should be understood as *intrinsic* to (or part of) an overall complex conscious

state. This stands in contrast to Rosenthal's standard HOT theory, where the HO state is *extrinsic* to (that is, entirely distinct from) its target mental state. The assumption about the extrinsic nature of the HOR has increasingly come under attack, and thus various hybrid representational theories can be found in the literature. Another motivation for this movement is renewed interest in a view somewhat closer to the one held by Franz Brentano (1874/1973) and various followers often associated with the phenomenological tradition.[2] To varying degrees, these hybrid views have in common the notion that conscious mental states represent *themselves* in some sense.

As was noted in the previous chapter, I have argued that when one has a first-order conscious state, the HOT is better viewed as *intrinsic* to the target state, so that we have a complex conscious state with parts (Gennaro 1996, 2006a). This is what I have called the wide intrinsicality view (WIV). Very briefly, we might say that conscious mental states should be understood (as Kant might have today) as combinations of passively received perceptual input and higher-order conceptual activity directed at that input. Higher-order concepts in metapsychological thoughts are presupposed in having first-order conscious states. I say much more about the WIV in chapter 4.

Another hybrid approach is advocated by Uriah Kriegel and is the subject of an entire anthology debating its merits (Kriegel and Williford 2006). Kriegel has used several different names for his "neo-Brentanian theory," such as the "same-order monitoring theory" and the "self-representational theory of consciousness." To be sure, the notion of a mental state representing itself or a mental state with one part representing another part needs further development. Nonetheless these authors agree that conscious mental states are, in some important sense, reflexive or self-directed. I criticize Kriegel's view in chapter 5.

Robert Van Gulick (2000, 2004, 2006) has also explored the alternative that the HO state is part of an overall conscious state. He calls such states "higher-order global states" (HOGS) whereby a lower-order unconscious state is "recruited" into a larger state, which becomes conscious partly due to the implicit self-awareness that one is in the lower-order state. Van Gulick has also suggested that conscious states can be understood materialistically as global brain states.

2.2 Defending Reductive Representationalism

2.2.1 Reduction and Explanation

Although it is possible to be a nonreductive representationalist (Chalmers 2004), most representational theories of consciousness are reductionist.

The classic notion at work is that consciousness, or individual conscious mental states, can be explained in terms of something *else* or in some other terms. It is worth mentioning that one prominent and influential model of reduction treats it as a form of *explanation* (Kemeny and Oppenheim 1956). Ney (2008) explains that "reductionists are those who take one theory or phenomenon to be reducible to some other theory or phenomenon. For example, a . . . reductionist about biological entities like cells might take such entities to be reducible to collections of physico-chemical entities like atoms and molecules." Explanation is certainly the ultimate goal of a reductionist theory of consciousness; that is, we want *to explain* what makes a mental state conscious.

Although Kemeny and Oppenheim had eliminativist leanings, one need not go that far in applying their model to consciousness. We can and should acknowledge that there really are conscious mental states, but also aspire to show that they can be explained in terms of a "base theory" devoid of consciousness-laden terms. Similarly, although their model of reduction employs the notion of reducing one *theory* to another, we can extend the idea to explaining *entities, events,* or *phenomena* such as conscious mental states. The familiar and successful example of explaining life in biological or cellular terms reminds us that such a reduction is not only possible but desirable.

Another reason to favor a reductionist approach is simply because nonreductive theories seem primarily motivated by the perceived lack of a plausible reductionist alternative. That is, it often seems to me that nonreductive accounts are mainly default positions stemming from the (correct or incorrect) conclusion that a given reductionist approach has failed. In some ways, antireductionism results from giving up on a reductionist approach. However, it would still seem odd to treat nonreductionism as an equally plausible explanation if there were also a viable reductionist account. And, of course, I view HOT theory as offering just such an account. It is hard to imagine that someone would adhere to a nonreductive approach just for its own sake. Are there, for example, any nonreductionists about life anymore?

With regard to explaining consciousness, however, we must distinguish between those who attempt such a reduction directly in *physicalistic*, such as neurophysiological, terms and those who do so using *mentalistic* terms, such as unconscious mental states or other cognitive notions. As I mentioned earlier, representationalists favor the latter strategy. I agree with Carruthers that those who currently attempt to reduce consciousness more directly in neural or physical terms "leap over too many explanatory levels

at once." (2005, 6). This is a point missed by Hardcastle (2004), for example, who mistakenly supposes that HOT theorists are chiefly motivated by the alleged nonreductionist divide between mind and brain or by some inherently mysterious explanatory gap (Levine 1983). Hardcastle also fails to appreciate that HOT theorists are very much open to a later second-step reduction to the neurophysiological, a point made by Rosenthal on several occasions.

Another general reason for a mentalistic approach is to blunt the force behind the so-called *multiple realizability* of conscious states. The idea here is that it seems perfectly possible for there to be other conscious beings, such as aliens or radically different animals, who can have those same kinds of mental states but be extremely different from us physiologically. It seems that commitment to a "type-type" identity theory, the view that mental state types (or properties) are identical with neural properties, leads to the undesirable result that only organisms with brains like ours can have conscious states (Fodor 1974). Thus most materialists wish to leave room for the possibility that mental properties can be "instantiated" in different kinds of organisms. Type-type identity theory is the very strong thesis that mental properties, such as "having a desire to drink some water" or "being in pain," are literally identical with a brain property of some kind. Such identities were originally meant to be understood as being on a par with, for example, the scientific identity between "being water" and "being composed of H_2O" (Place 1956; Smart 1959), but this failed to acknowledge the multiple realizability of mental states. So I take it that one advantage of HOT theory is that it is not committed to any *direct* reduction of consciousness to neural activity. Nonetheless HOT theorists are typically still materialists who desire to show how HOT theory might be realized in *our* brains.

2.2.2 Gaps, Zombies, and Phenomenal Concepts

Some philosophers have argued that there is a potentially permanent explanatory gap between our understanding of consciousness and the physical world (Levine 1983, 2001) and that we do not, or even *cannot*, understand how consciousness arises from brain activity (Chalmers 1995). If they are correct, then there could not be an ultimately successful reductionist account of consciousness.

McGinn (1991), for example, goes so far as to argue that we are not cognitively equipped to understand how consciousness is produced by the brain. We are "cognitively closed" with respect to the mind–body problem much as a rat or dog is cognitively incapable of solving, or even understanding, calculus problems. McGinn concedes that some brain property

produces conscious experience, but we cannot understand how it does so, and we cannot come to know what that brain property is. Our concept-forming mechanisms will not allow us to grasp the physical and causal basis of consciousness. McGinn does not entirely rest his argument on past failed attempts at explaining consciousness in physical terms. Instead he presents a distinct argument for his pessimistic conclusion. McGinn observes that we do not have a mental faculty that can access both consciousness and the brain. We access consciousness through introspection, but our access to the brain comes through outer spatial senses. Thus we have no way to access both the brain and consciousness together, and therefore any explanatory link between them is forever beyond our reach.

Finally, an appeal to the possibility of zombies is also sometimes taken both as a problem for materialism and as a more positive argument for some form of dualism, such as property dualism. The philosophical notion of a "zombie" refers to conceivable creatures that are physically indistinguishable from us but lack consciousness entirely (Chalmers 1996). It certainly seems logically possible for such creatures to exist: "The conceivability of zombies seems . . . obvious to me. . . . While this possibility is probably empirically impossible, it certainly seems that a coherent situation is described; I can discern no contradiction in the description" (Chalmers 1996, 96). Philosophers often contrast what is logically possible (in the sense of "that which is not self-contradictory") from what is empirically possible given the actual laws of nature. Thus it is logically possible for me to jump fifty feet in the air, but not empirically possible. The objection, then, typically proceeds from such a possibility to the conclusion that materialism is false because it would seem to rule out that possibility. It has been fairly widely accepted (since Kripke 1972) that all identity statements are necessarily true (that is, true in all "possible worlds"), and the same should therefore hold for mind–brain identity claims. Since the possibility of zombies shows that mind–brain identity claims do not, then we should conclude that materialism is false.

Some philosophers explicitly draw antimaterialist and antireductionist conclusions from these considerations (Chalmers 1996), while others do not view them as a threat to the *metaphysics* of materialism (McGinn 1991; Levine 2001). Either way, however, I think there is a plethora of plausible replies to the foregoing lines of argument that would take me too far afield from my main topic.[3]

I do, however, wish to pause to address one influential reply that involves a claim about a special class of concepts called *phenomenal concepts* (Loar 1990, 1997). Phenomenal concepts are *recognitional* concepts. To have

the phenomenal concept of blueness is to be able to recognize experiences of blueness while having them. The recognitional concept of blueness refers *directly* to its referent (the physical property of blueness), so there is no other property involved in the reference fixing. Phenomenal concepts are *indexical or demonstrative concepts* applied to phenomenal states via introspection (Lycan 1996). Carruthers, for example, describes purely recognitional concepts as those "we either have, or can form . . . that lack any conceptual connections with other concepts of ours, whether physical, functional, or intentional. I can, as it were, just recognize a given type of experience as *this* each time it occurs, where my concept *this* lacks any conceptual connections with any other concepts of mine—even the concept *experience*" (2005, 67).

According to Loar, Carruthers, and others, these concepts mislead us into thinking that any alleged explanatory gap is deeper and more troublesome than it really is. Ironically, it is perhaps McGinn's own observation about our two distinct concept-forming mechanisms that is used to blunt the force of the problems just described. Given our possession of phenomenal concepts, Loar and others reply that any alleged explanatory gap or lack of identity between the mental and physical can be explained away. If we possess purely recognitional concepts of the form "*This* type of experience," we will always be able to have that thought while, at the same time, conceiving of the absence of any corresponding physical or intentional property. On the one side, we are using scientific third-person concepts, and on the other, we are employing phenomenal concepts. We are, perhaps, simply not in a position to understand completely the connection between the two, but the mere possibility of, say, zombies is explained away in a manner that is harmless to materialism. It may be that there is a good reason why such zombie scenarios seem possible, namely, that we do not (at least not yet) see what the necessary connection is between neural events and conscious mental events.[4]

For my own part, I am not quite convinced that there are phenomenal concepts, at least in the way they are often defined. First, it is unclear that HO theorists need to invoke them to provide a reductionist account of consciousness in mentalistic terms. The so-called phenomenal concept strategy is primarily used by those who wish to reduce consciousness to something expressed in overtly physical terms. As we have seen, this is not the strategy of a HO theorist.

Second, it is not clear to me that there are any concepts that have *no* "conceptual connections with other concepts, whether physical, functional, or intentional," as Carruthers puts it. It seems to me that even such

alleged recognitional or indexical concepts have at least some relation to other concepts possessed by the subject even if they are not concepts framed in physicalistic terms. Rosenthal shares my skepticism: "Even when we recognize something without knowing what type of thing it is, we always can say something about it" (2005, 207). At minimum, there would seem to be many comparative concepts involved in any such description, such as when one sees a darker or lighter shade of a color than has been seen up to that point.

Third, I suppose that one *could* think of HOTs as indexical or demonstrative thoughts and thus akin to phenomenal concepts in this respect. The idea would be to think of HOTs as having the form "*I* am in *this* mental state" or "*This* is the mental state *I* am in," since "I" and "this" are demonstratives and indexicals.[5] But I fail to see the advantage of this approach over standard HOTs of the form "I am in mental state M." Perhaps "I am in *this* mental state" is less conceptually sophisticated, which might help with respect to the Animals and Infants Theses, but there are still the concepts "I" and "mental state" as constituents of those thoughts. Moreover, I take the fact that there are concepts in the HOTs to be an *advantage* of HOT theory over, say, HOP theory, for reasons we will see in later chapters.

Perhaps most important for those who do advocate reductionism in purely physical terms, however, is simply recognizing that different concepts can pick out the same property or object in the world. Out in the world there is only the one "stuff," which we can conceptualize either as "water" or as "H_2O." Recall again the Fregean distinction between meaning (or "sense") and reference. Two concepts can have different meanings but refer to the same property or object, much like "Venus" and "the Morning Star." Materialists, then, explain that it is essential to distinguish between mental properties and our concepts of those properties. By analogy, there are phenomenal concepts that employ a phenomenal property to refer to some conscious mental states, such as a sensation of red. In contrast, we can also use concepts couched in physical or neurophysiological terms to refer to that same mental state from the third-person point of view. There is thus only one conscious mental state conceptualized in two different ways: either by employing first-person experiential phenomenal concepts or by employing third-person neurophysiological concepts. It may then just be a "brute fact" about the world that there are such identities, and the appearance of arbitrariness between brain properties and mental properties is just that—an *apparent* problem leading many to wonder about the alleged explanatory gap. Qualia could then, after all, be identical to physical properties. Moreover, this response provides a diagnosis for *why there even*

seems to be such a gap, namely, that we use very different concepts to pick out the same property. With respect to the more general issue of reduction, however, I think that Carruthers (2005, chap. 2) and Block and Stalnaker (1999) rightly criticize the notion that a priori conditionals between the physical and mental are *required* for a successful reduction, at least for most standard models of explanation (Chalmers and Jackson 2001). I return to this matter in chapter 4.

In any case, I think it is best to adopt what we might call *methodological reductionism*, whereby we attempt, as a matter of strategy or method, to reduce consciousness to intentionality (or something cognitive) unless it is clearly impossible. It is not time to give up. How can success for such a strategy be ruled out a priori or so soon? It seems premature to declare that any kind of successful reduction is forever hopeless. Of course, there are philosophers who believe more specifically that intentionality itself entails or involves consciousness, which would then make such a reduction impossible. It is to this issue that I now turn.

2.3 Consciousness and Intentionality

The relationship between intentionality and consciousness is itself a major ongoing area of dispute, with some arguing that genuine intentionality actually presupposes consciousness in some way (Searle 1992; Siewart 1998; Horgan and Tienson 2002; Pitt 2004; Georgalis 2006). One way to frame the issue is in terms of the question "Does mentality entail consciousness? (Gennaro 1995). Notice that an affirmative answer results in a very strong claim; that is, having intentional states (such as beliefs, thoughts, and desires) *entails* having conscious states. I argue that this is much too strong.

2.3.1 Searle's Connection Principle

It will be useful first to critically examine Searle's well-known and controversial Connection Principle (1992, 132) in support of the entailment claim. It says:

(CP) Every unconscious intentional state is at least potentially conscious.

Searle similarly tells us that the "notion of an unconscious mental state implies accessibility to consciousness" (152). Much of Searle's argument for CP rests on the notion that every intentional state has "aspectual shape," which can ultimately be accounted for only via consciousness. The idea is that genuine intentional content must ultimately "seem" a certain way to a creature and so presumably involves a conscious first-person point of view.

This is largely because Searle thinks that this is the only way to account for the *intensionality* (with an *s*) of intentional states. For example, if a person P has the (unconscious) belief that there is water in the pool, P must be able to conceive of that substance under the aspect of "water" (as opposed to, say, H_2O). But since only conscious intentionality is *intrinsically* aspectual, the idea of an unconscious intentional state is parasitic on the conscious variety.

It is indeed widely accepted that intensionality is a mark of intentional states. The idea is that substituting co-referring terms in a statement does not necessarily preserve truth value. A four-year-old child (who knows nothing about chemistry) can know or believe that there is water in the pool, but it would be false to say that she knows or believes that there is H_2O in the pool. Searle's claim, however, is that for there to be unconscious aspectual shape, it must be possible for the organism to have intrinsic aspectual shape. And intrinsic aspectual shape can only arise with reference to a conscious point of view. So what distinguishes an unconscious *mental* state from other neural happenings is that it is potentially conscious.

Nonetheless, numerous decisive objections to CP have been raised over the years.[6] I review some here.

First, the notion of "potential" at work in CP must obviously not be a logical or metaphysical possibility. That would surely be too strong. Thus nomologically possible or psychologically possible seems much more reasonable. But then if we take CP literally, Searle faces the problem that it mistakenly rules out a host of abnormal psychological phenomena, such as deeply repressed states or any unconscious state that could not *in fact* become conscious owing to brain lesions and the like (Rosenthal 1990).

Second, there seems to be no way for CP to acknowledge intentional states that occur via some forms of perceptual processing. For example, there would seem to be two visual pathways in the brain (Milner and Goodale 1995). Visual processing along the *ventral stream* pathway is conscious. But visual processing also occurs along the *dorsal stream* visual pathway, which generates representations not accessible to consciousness. The dorsal stream functions more like an unconscious (and very fast) visual motor system that causes the relevant behavior due to systematic tracking relations with the environment. One might deny that dorsal-stream representations are genuinely intentional, but this would be an extremely odd line to take.

Third, CP seems to entail what Shani calls a "denial of gradualism," whereby converging lines of empirical evidence show that "the evolution of subjectivity is a gradual process manifesting various levels of ascending

complexity, each serving as a platform for the emergence of . . . subjective existence" (2007, 59; see also Shani 2008). As is evidenced by the previous objection, perhaps there are lower animals (such as lizards and rodents) that *only* have the dorsal-stream visual processing. This seems likely on at least *some* level of evolutionary development. I fail to see any reason, however, to hold that such animals cannot have any genuinely contentful intentional states (including perceptual states) unless those states could also be conscious. At the least, it seems *possible* for such an organism to exist. We can and should allow for degrees of intentionality and understanding of the environment.

Fourth, another way to approach the matter is by answering the following question: Can significant explanatory power be achieved by making intentional attributions without attributions of consciousness? It seems to me that the answer is clearly yes, as the animals' case in the previous paragraph shows. We would, I suggest, still rightly attribute all unconscious intentional states to such animals. Would or should we withdraw intentional attributions to an animal if we later come to agree that it is not conscious? I don't think so. Such attributions are useful in explaining and predicting animal behavior, but it does not follow that they have merely "as-if" intentionality. In some cases, we may not know if they are conscious. The same, I suggest, would hold for advanced robots. This is not necessarily to embrace some kind of antirealist Dennettean "intentional stance" position (Dennett 1987). For one thing, we might still agree that those systems have genuine internal mental representations.

Finally, the foregoing considerations show us how to challenge more directly Searle's central premise that there cannot be intrinsic unconscious aspectual shape. Searle thinks that genuine cases of aspectual shape and intensionality cannot be revealed from mere third-person evidence (behavioral or otherwise). For example, he would presumably hold that no third-person evidence could ever justify an attribution of a belief about water as opposed to a belief about H_2O. But surely a counterexample is possible. For example, if an unconscious robot displays enough sophisticated behavior that it systematically locates and recognizes a bottle labeled "water" as opposed to bottles labeled "H_2O" (among many other water-related behaviors), then we may be warranted in attributing to it the former belief (that is, the belief about where the bottle of water is). Even Searle recognizes that one can have, say, a desire for water and not have a desire for H_2O, though water and H_2O are the same. His mistake, however, is to suppose that nothing short of a first-person subjective point of view can justify the attribution of one state but not the other (Van Gulick 1995a,b).

To be fair, however, Searle's line of argument does raise a genuine challenge for all naturalistic (or reductionist) theories of mental content, namely, just how to specify or determine intentional contents without a first-person or subjective point of view. One problem raised by Searle is that third-person evidence always leaves the aspectual shape underdetermined to some extent (Searle 1992, 158, 163–164). Or, as Quine (1960) might put it, there would be indeterminacy of intentional content without the first-person evidence.

Several replies are in order here. (1) If the above robot-bottle story makes any sense at all, it is not clear that all such intentional content must be undetermined or underdetermined. Under certain conditions, it at least seems possible to attribute all unconscious intentional states to a system. (2) In some ways, then, Searle simply begs the question against naturalistic theories of content. He is right to demand that his opponent offer a workable theory along these lines, but to rule out success up front again seems premature. Moreover, some of us are not entirely uncomfortable with a theory of content that allows for some degree of indeterminacy if it has other theoretical advantages. (3) Searle seems to think that determinacy can be gained in a straightforward way once we include the first-person point of view. But is this so obvious? The real force behind Quine's position, I take it, is that even the first-person point of view does not always fix what we mean by a term or concept. It is not always obvious just what *I* mean by "water" or "rabbit." Introspective evidence, while important and often reliable, is not infallible and does not always lead to determinacy of content. Does such evidence really tell me whether or not I mean "undetached rabbit parts" when I think about a "rabbit"?

Another important question can be put as follows: what makes a state a *mental* state (as opposed to, say, a mere information-carrying state)? This question can surely be answered without invoking consciousness at all. One option is to hold that the creature in question must have complex-enough behavior such that simple mechanistic explanations are not sufficient to explain its behavior. More positively, we might demand that creatures or systems display a significant degree of inferential integration (or "promiscuity") among their intentional states (Stich 1978). The contents of, say, beliefs and desires are interconnected in various ways; thus, beliefs and desires acquire their content within a web or network of beliefs. So, for example, the more "informationally encapsulated" a state is (Fodor 1983), such as in early visual processing, the less likely it is to count as a mental state.

These considerations can also be used in response to the slippery-slope argument that any attempt to explain intentionality that detaches it from

consciousness leads to the absurd conclusion that intentionality would then be everywhere (Searle 1992, 1995; Strawson 2004). Stomachs would have mental lives, and water really *tries* (that is, "desires") to get to the bottom of the hill. Once again, these absurd implications can be blocked by recognizing that stomachs and rivers do not meet the criterion above, namely, that there is no significant degree of inferential connections among their states. Moreover, attributing intentionality to stomachs and rivers does not add any explanatory value to a purely mechanistic (or informational) account. In conclusion, then, I think that CP is false.

Of course, the general claim that "mentality entails consciousness" remains ambiguous. There are numerous interpretations depending on which kinds of mental states are at issue, as well as whether or not we are concerned with state or creature consciousness.[7] I think most interpretations are false, but let us briefly consider the following two:

(1) A creature or system cannot have all unconscious beliefs and desires (or "goals").

(2) A creature or system cannot have all unconscious pains, frustrations, or sufferings.

As I have argued, I think that (1) is false, but (2) might very well be true. For (1), the *system or creature* might be utterly unconscious but have such intentional states, whereas in (2) a creature would arguably have to be conscious to have any genuine pains or sufferings. Perhaps the difference lies in the fact that some intentional states, such as beliefs, are best understood as dispositions to behave in various ways. On the other hand, (2) does seem true to me. It at least seems much more reasonable to claim that even if there are *individual* unconscious pains and (perhaps) frustrations, we would likely not attribute such states to a creature if we believed that it was not conscious at all. It seems odd to talk about the frustrations, sufferings, or pains of an utterly unconscious creature or robot. The reason for this is perhaps that our very concept of "suffering" or "pain" is more closely tied to consciousness. Unlike Searle, however, the connection here is not one of state consciousness but rather one of overall creature consciousness. That is, for example, I hold not that *each* individual pain must be potentially conscious but that attributions of unconscious pains make sense only if we also think that the *creature* in question is conscious.

2.3.2 Phenomenal Intentionality

Reductive representationalists hold that intentionality is separable from consciousness, a view that Horgan and Tienson (2002) reject and call

separatism. They argue for what is called *phenomenal intentionality* or "cognitive phenomenology." One rationale for separatism is to make a reductionist explanation of consciousness possible. But if intentionality is *deeply* intertwined with consciousness, then a reductionist explanation would be difficult or perhaps even impossible to obtain. And some argue that beliefs, desires, and other intentional states themselves have phenomenology.

Horgan and Tienson distinguish the Intentionality of Phenomenology (IP) from the Phenomenology of Intentionality (PI). They state PI as follows:

(PI) "Mental states of the sort commonly cited as *paradigmatically intentional* . . . *when conscious*, have phenomenal character that is inseparable from their intentional content" (2002, 520; italics mine).

In addition they advocate the claim that "there is a *kind* of intentionality, pervasive in *human* mental life, that is constitutively determined by phenomenology alone" (520; italics mine).

Although Horgan and Tienson's purpose is not explicitly to reject reductive representationalism, the impression given is that PI is a threat to reductionism or naturalism. However, a careful reading of the foregoing quotations reveals that PI is compatible with reductionism and consistent with a negative answer to the question "Does mentality entail consciousness?" The main issue, as I see it, is their starting point, namely, the first-person *human* point of view. They primarily have in mind *paradigmatic human* cases of intentional states, which they argue involve phenomenology. So, for example, there is something it is like for us to *think* that rabbits have tails, *believe* that ten plus ten equals twenty, or *desire* Indian food. The consciousness in question is presumably not merely *accompanying* associated images of rabbits or food (Lormand 1996) but rather intrinsic to the intentional states themselves. But it still does not follow that intentionality per se entails consciousness or phenomenology, as we have already seen in the previous subsection. There may be some intentional states that could not become conscious or even an organism (or robot) with all unconscious intentional states.[8]

Moreover, Horgan and Tienson often seem more concerned with the viability of narrow content than with the separability of intentionality and consciousness. But as far as I can see, believing in narrow content is also not inconsistent with reductionism (Carruthers 2000, 2005). Like Carruthers, I also hold that there is narrow content. This combination of views may not be *typical* among representationalists, but it is hardly inconsistent.

We should also distinguish, as Horgan and Tienson do, the phenomenology of *attitude type* (desires, thoughts, beliefs, wonderings, etc.) from the

phenomenology of *content* (the same attitude but with different content). I raise three points here:

(1) I am inclined to agree that there is phenomenal intentionality for most intentional attitude types. It does indeed seem right to hold that there is something it is like to think that rabbits have tails, believe that ten plus ten equals twenty, or have a desire for some Indian food. But, again, this is no threat to reductionism, because a representationalist can simply agree that those kinds of mental states need to be added to the list of conscious mental states for which we need an explanation. For example, a HOT theorist might accept that one's thoughts or hopes become conscious when a suitable (unconscious) HOT is directed at it. There is little reason to resist the idea that my (conscious) desire to write a good book or my (conscious) thought that I am on sabbatical has a phenomenological aspect. But this does not imply that each individual intentional state is actually or potentially conscious.

(2) It seems to me, however, that there is something importantly different about beliefs and knowledge, on the one hand, and desires, wonderings, and thoughts, on the other. Beliefs and knowledge seem to be purely dispositional states, in contrast to, say, occurrent episodes of thinking. In the former case, I think what we really have in mind are cases of consciously *introspecting* our beliefs or knowledge so that the *objects* of conscious thoughts are conscious. Is there something it is like to believe, *as opposed to* think about, the cat in the tree? I don't think so. Thus it is not even clear that there are first-order *conscious* beliefs or knowledge at all (Gennaro 1996, 36–43).

(3) It is also doubtful that there is a different phenomenology for *every* change in *content*. For example, let us agree that there is a phenomenological difference between thinking about a one-thousand-sided figure and thinking about a four-sided figure. But it still seems wrong to hold that there is a phenomenological difference between thinking about a 999-sided figure and a 998-sided figure. Is there a phenomenological difference between wondering whether a distant star is 800 light-years away or 850 light-years away? Just how fine grained can contents be such that there is a phenomenological difference? One can easily generate an infinite number of different contents for each single attitude type, but it seems unlikely that there is a phenomenological difference for each pair.

Finally, it is worth remembering that in HOT theory (or something close to it), consciousness entails intentionality, but not vice versa. However, an appropriate representation *of a* representation does entail consciousness and is constitutive of it. I now turn to a preliminary defense of HOT theory.

2.4 HOT Theory: An Initial Defense

In this section, I offer a preliminary defense of HOT theory. I ask the reader for some patience as a more thorough defense and additional details of my own theory will become clearer throughout the book.

2.4.1 The Transitivity Principle

It is natural to start with the highly intuitive claim that has come to be known as the Transitivity Principle (TP). One motivation for HOT theory is the desire to use this principle to explain what differentiates conscious and unconscious mental states:

(TP) A conscious state is a state whose subject is, in some way, aware of being in it (Rosenthal 2000a, 2005).[9]

Thus, when one has a conscious state, one is aware of being in that state. For example, if I am having a conscious desire or pain, I am aware of having that desire or pain. HOT theory says that the HOT is of the form "I am in M now," where M references a mental state. Conversely, the idea that I could be having a conscious state while totally *unaware* of being in that state seems very odd (if not an outright contradiction). A mental state of which the subject is completely unaware is clearly an *un*conscious state. For example, I would not be aware of having a subliminal perception, and thus it is an unconscious perception. I view the TP primarily as an a priori or conceptual truth about the nature of conscious states. It is interesting to note that many non-HOT theorists agree with the TP, especially those who endorse some form of self-representationalism according to which conscious mental states are also directed back at themselves in some sense.[10]

One can also find a similar claim in Lycan's (2001a) argument where premise (1) just is the TP. Moreover, he treats it as a "definition," which suggests that it is a conceptual truth. The entire argument runs as follows:

(1) A conscious state is a mental state whose subject is aware of being in it.
(2) The "of" in (1) is the "of" of intentionality; what one is aware of is an intentional object of the awareness.
(3) Intentionality is representational; a state has a thing as its intentional object only if it represents that thing.
Therefore,
(4) Awareness of a mental state is a representation of that state. (From 2, 3)
Therefore,
(5) A conscious state is a state that is itself represented by another of the subject's mental states. (1, 4)

I should say that Lycan's argument does not necessarily support HOT theory as opposed to his favored HOP theory, but I will argue against HOP theory in the next chapter. Moreover, the argument does not, strictly speaking, rule out a self-representational account because (5) does not necessarily follow from (1) and (4). For example, a self-representationalist will say that the representing state need not be *distinct* from the represented state (Gerken 2008). To be fair to Lycan, however, much of the work on self-representationalism referenced in this book occurred after his 2001a piece was published. In addition, Lycan clearly intended to be arguing for a *reductive* representational account, which is typically not the self-representational view. Thus Lycan's argument might be too simple, but it can be supplemented by additional argumentation. HOT theorists often employ an "argument by elimination" strategy against various other theories of consciousness (Carruthers 2000; Rosenthal 2004).

One might object that many HO theorists hold that the TP is an *empirical* (as opposed to an a priori) claim. Indeed, Rosenthal himself says, "The theory doesn't appeal to, nor is it intended to reflect, any conceptual or metaphysically necessary truths" (2005, 9). But he also refers to the TP as a "truism" (8), which seems to suggest that it is a conceptual, or at least "folk psychological," truth of some kind. Rosenthal also often asserts the "intuitively obvious" truth of TP and seems to use a priori reasoning in various places. Bill Lycan has also told me, in e-mail correspondence, that he wonders if HO theories are "nearly trivially true." In any case, if I differ from other HO theorists on the extent to which HO theory is a conceptual truth or is known a priori, then so be it.

There is also an importantly related issue here. If "an empirical claim" means "in principle empirically falsifiable" or "consistent with and sometimes supported by empirical and scientific evidence," then I certainly agree that HO theory is empirical. A conceptual or necessary truth might *also* be empirical in the sense that it can sometimes also be supported or falsified by empirical evidence. We might *claim* to know that some proposition is true a priori but then come across empirical findings that *falsify* it. Indeed, this happens often in philosophy of mind when facts about abnormal psychological phenomena call into question what seem to be obvious conceptual truths, such as when the existence of Anton's syndrome (blindness denial) forces us to doubt the view that we cannot be mistaken about our ability to see. Another case would be falsifying what Descartes surely took to be conceptually true, namely, a kind of "self-intimation" thesis that denies the very possibility of unconscious mental states and says that if one has a mental state, then one knows that one is in it. In such cases, we typically later conclude that these propositions were not really known in the first place.

2.4.2 Other Aspects of HOT Theory

Another central motivation for HOT theory is that it purports to help explain how the acquisition and application of concepts can transform our phenomenological experience. Rosenthal invokes this idea with the help of several well-known examples (2005, 187–188). For example, acquiring various concepts from a wine-tasting course will lead to different experiences from those enjoyed before the course. I acquire more fine-grained wine-related concepts, such as "dry" and "heavy," which in turn can figure into my HOTs and thus alter my conscious experiences. As is widely held, I will literally have different qualia due to the change in my conceptual repertoire. As we learn more concepts, we have more fine-grained experiences and thus experience more qualitative complexities. Conversely, those with a more limited conceptual repertoire, such as infants and animals, will have a more coarse-grained set of experiences. Much the same goes for other sensory modalities, such as the way that I experience a painting after learning more about artwork and color. These considerations do not, of course, by themselves prove that newly acquired concepts are *constitutive* parts of the resulting conscious states, as opposed merely to having a *causal* impact on those states. Nonetheless, I will argue in subsequent chapters that it is more plausible to suppose that concepts are indeed constitutive parts of conscious states because it is better to construe (unconscious) HOTs as intimately bound up with the lower-order states.

Let us also consider a common initial objection to HOR theories, namely, that they are circular and lead to an infinite regress. For example, it might seem that HOT theory results in circularity by defining consciousness in terms of HOTs. It might also seem that an infinite regress results because a conscious mental state must be accompanied by a HOT, which must in turn be accompanied by another HOT, ad infinitum. However, the standard reply is that when a conscious mental state is a first-order world-directed state, the HOT is *not* itself conscious; otherwise circularity and an infinite regress would follow. When the HOT is itself conscious, there is a yet-higher-order (or third-order) thought directed at the second-order state. In this case, we have *introspection*, which involves a conscious HOT directed at an inner mental state. When one introspects, one's attention is directed back into one's mind. For example, what makes my desire to write a good book a conscious *first-order* desire is that an unconscious HOT is directed at the desire. In this case, my conscious focus is directed at the book and my computer screen, so I am not consciously aware of having the HOT from the first-person point of view. When I introspect that desire, however, I then have a *conscious* HOT (accompanied by a yet higher, third-order, HOT) directed at the desire itself (Rosenthal 1986, 1997). Figure 2.1 is one way to illustrate HOT theory.

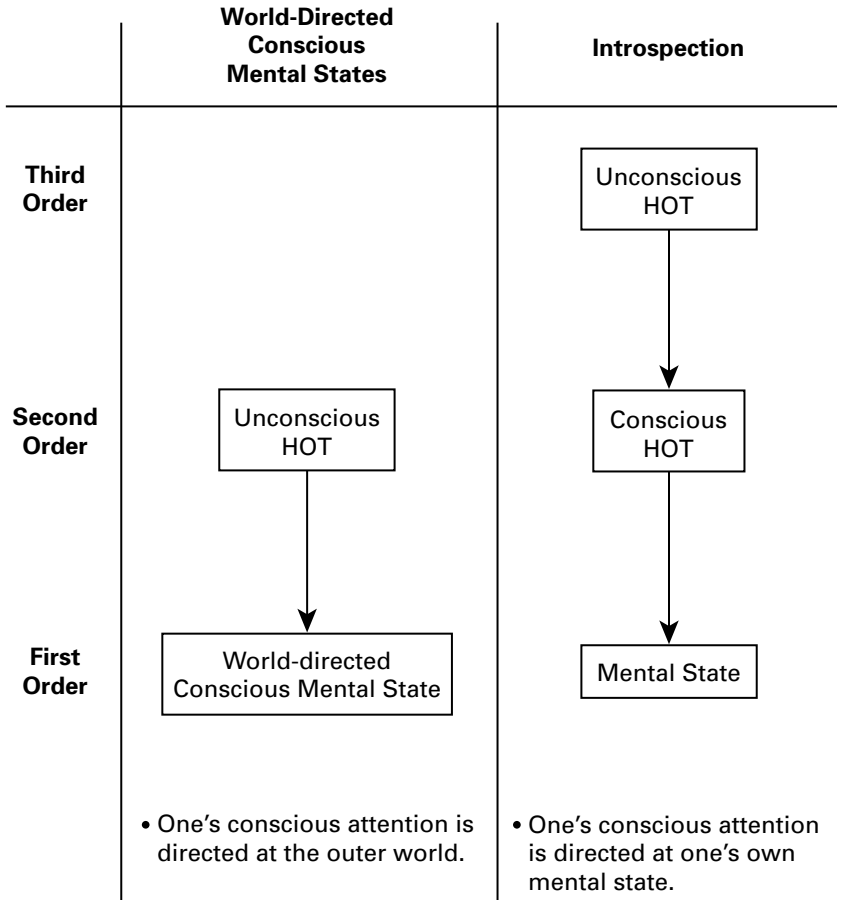


Figure 2.1
The structure of conscious mental states according to the HOT theory of consciousness.

Another related and compelling rationale for HOT theory and the TP is as follows (based on Rosenthal 2004, 24): A non-HOT theorist might still agree with HOT theory as an account of *introspection* or *reflection*, namely, that it involves a conscious thought about a mental state (Block 1995). This seems to be a fairly common sense definition of introspection that includes the notion that introspection involves conceptual activity. It also seems reasonable for anyone to hold that when a mental state is unconscious, there is no HOT at all. But then it stands to reason that there should be something "in between" those two cases, that is, when one has a first-order conscious state. So what is in between no HOT at all and a conscious HOT? The answer, of course, is an unconscious HOT, which is precisely what HOT theory says. Moreover, this explains what happens when there is a transition from a first-order conscious state to an introspective state: an unconscious HOT becomes conscious.[11]

HO theorists further agree that the HO state must become aware of the LO state noninferentially. We might even suppose, say, that the HO state must be caused noninferentially by the LO state to make it conscious. The point of this condition is mainly to rule out alleged counterexamples to HO theory, such as cases where I become aware of my unconscious desire to kill my boss because I have consciously inferred it from a session with a psychiatrist, or where my envy becomes conscious after making inferences based on my own behavior. The characteristic *feel* of such a conscious desire or envy may be absent in these cases, but since awareness of them arose via conscious inference, the HO theorist accounts for them by adding this noninferential condition.

Finally, it is worth mentioning that there is no reason in principle to rule out the possibility of experimental data supporting HOT theory and, in particular, the continuous presence of unconscious HOTs. Despite her scathing but somewhat misdirected criticism of HOT theory, Hardcastle (2004, 290–294) suggests that the ubiquitous presence of unconscious HOTs could find empirical support via a modified priming task. There is no reason why some of the methods used to indicate the presence of unconscious *first-order* mental states could not, if suitably modified, also be used to indicate the presence of unconscious HOTs. For example, one well-known method is known as *subliminal priming*, which refers to the effects on subsequent behavior of stimuli that are not consciously detected (Marcel 1983). Unconscious mental processes can influence our conscious mental states.

For example, Jacoby, Lindsay, and Toth (1992) briefly presented completed words before presenting a target word stem, such as presenting RESPOND followed by ___OND. But then subjects were told *not* to use the

completed word in suggesting that it would complete the stem. Subjects would also be primed unconsciously to give the flashed word although they were instructed to disregard it. In such an opposition condition, subjects would take longer to answer questions for which they had just been primed with an answer that they could not use. But when they were told to use the completed word, priming would work to their advantage. Their reaction times should be shorter. By comparing response times between these two conditions, as well as their respective error rates, we get some idea of the influence that unconscious states can have on their conscious answers.

Hardcastle suggests that we "can and should use a similar methodology to determine whether we have unconscious HOTs . . . co-active with any conscious states. . . . We need a priming task that would test whether we can recognize that we were aware of a series of target conscious events faster or with fewer errors than other aspects of the same events. If we can, then that would be some evidence that we are unconsciously aware that we are aware" (2004, 292). She gives an example of one possible experiment. We flash a series of simple scenes (such as a cat on a mat or a dog with a bone) for a half second or so, long enough to reach consciousness. Each scene is then replaced by the same masking stimulus, which prevents subjects from studying the stimulus. We can then ask about their conscious experience (did you see a bone?) or about the scene (was the dog next to the bone?). With appropriate controls in place, if we have unconscious HOTs "accompanying all conscious experiences, then HOTs should prime our behavior with regard to reacting to the fact that we are conscious" (292), and we should answer the former questions (about conscious experience) with fewer errors than the latter (about the scene). To my knowledge, however, these kinds of experiments have not been done to date. Aside from this specific suggestion, there should be some way to design experiments that could serve as experimental evidence for or against HOT theory.

2.5 More on Mental Content

Now that we have established a prima facie case for HOT theory, let us return to mental content. In the end, I think that a HOT theorist *could be* relatively neutral with respect to theories of mental content. It is not clear that a HOT theorist *must* be wedded to any particular theory of content. Nonetheless it is fair to ask any proponent of HOT theory just where one's sympathies lie, and some of the details might also affect how one handles various objections. Three areas need to be addressed:

(1) The first has to do with exactly which theory of content is preferred. That is, just how do mental representations and their constituent concepts acquire their content (or "meaning")? What determines their content? Many such theories are on offer. Perhaps the most common division among naturalistic views is between *causal-informational* (Stampe 1977; Dretske 1981, 1988; Fodor 1981, 1990) and *functional* theories (Block 1986; Harman 1973). Causal-informational theories hold that the content of a mental representation is grounded in the information it carries about what does, or would, cause it to occur. Mental states acquire their content by standing in appropriate causal relations to objects and properties in the world. The basic idea is that, say, thoughts about dogs are about dogs, and *mean* "dog," because dogs cause the thoughts that our minds use to keep track of dogs. Functional theories hold that the content of a mental representation is grounded in its causal or inferential relations to other mental representations. My preference is with causal theories, though they are also sometimes supplemented in various ways, such as by teleological or biological considerations.[12]

Causal theories, however, do face some well-known difficulties. For example, a very crude causal theory cannot be sufficient for specifically *mental* content, not to mention *conscious* content. For one thing, causal relations abound where no mentality exists at all, such as with tree rings and thermostats. Perhaps most important is the *disjunction problem*, which shows that a simple causal story cannot properly isolate the correct causal relation. A horse *might* normally cause the mental tokening of the concept "horse," but why not "saddle" instead? We thus encounter the related possibility and problem of *misrepresentation*, which any theory of representation should recognize. Perhaps cows (say, not seen in proper lighting) sometimes cause mental representations of "horse." How is this explained? Does "horse," then, represent *either* cows *or* horses? Getting the extension of a mental representation right is paramount for any theory of content. It should be noted that we are mainly concerned with empirical objects and properties.

I will not pretend to have a novel solution to these ongoing disputes. Clever attempts to solve these problems from the likes of Dretske and Fodor have left many dissatisfied. For example, Dretske posits a learning period during which mental content is fixed. Once the learning period ends, it is then possible for the mental representation to be misapplied to (and thus to misrepresent) the corresponding object or property. Although this overall strategy may be right in some regard, it has been met with significant criticism (Slater 1994; Prinz 2002). For example, it is well known that children overgeneralize their concepts during the learning process itself.

Fodor (1987, 1990) puts forth an *asymmetric dependence* theory based on the observation that informational relations depend on representational relations, but not vice versa. An important asymmetry is at work here. For example, if mental representations (or tokens of a mental state type) are reliably caused by horses *and* cows-on-dark-nights, then they also carry information about all those objects. If, however, the mental representation "horse" is tokened in response to a cow on a dark night, this tokening depends on the more fundamental relation between horses and horse representations. In other words, if it were not the case that horses caused "horse" concepts or mental representations, then cows would not token "horse" either. Thus the content-determining causes are more fundamental in an important sense.

For my money, the best attempts to handle these problems can be found in the related work of Rupert (1999) and Prinz (2002). They build on Dretske's notion of a learning period but appeal to the *actual history* of causal interactions between a mental representation and what it represents. Rupert offers a modified causal view, at least for natural kind terms, called the *best test* or *causal-developmental* theory, according to which there is an actual history requirement for a mental representation to acquire its content accurately. The basic idea is that content is determined by a substantive developmental process shaped by a subject's developmental interaction with the environment. A mental representation R "has as its extension the members of natural kind K if and only if members of K are *more efficient* in their causing of [R] in S than are the members of any other natural kind" (Rupert 1999, 323; italics mine). The notion of "efficiency" is cashed out in terms of numerical comparisons between the *past relative frequencies* (PRFs) of certain causal interactions (cf. Usher 2001).

So in response to the disjunction problem, the idea is that although every cat is a mammal, the PRF of cats relative to the concept "cat" is much higher than mammals relative to that concept. Only PRFs resulting from a substantial number of interactions matter. With respect to the earlier example, the concept "horse" will not represent cows because that concept will be caused much more frequently by horses. It is the *success rates* (that is, the percentage) of object or property to mental representation R that determine content, not necessarily the most common stimulant of R. Similar considerations explain misrepresentation after a concept is acquired.

In a somewhat related manner, Prinz (2002, 250) urges that the "intentional content of a concept is the class of things to which the object(s) that caused the original creation of that concept belong." Again, what matters

is the actual causal history of a concept. More specifically, mental content is "identified with those things that *actually* caused the first tokenings of a concept (what I call the 'incipient causes'), not what *would* have caused them" (250). So *both* nomological covariance *and* incipient causes are necessary to determine intentional content. "Incipient causes are a special subset of actual causes" (251). Prinz (2002, 251) summarizes as follows:

X is the *intentional content* of concept C if (a) Xs nomologically covary with tokens of C *and* (b) an X was the incipient cause of C.

Prinz explains that clause (b) can solve the disjunction problem. Horses, not cows, are the basis on which the concept "horse" is formed. Not just any causes that happen to occur in the actual history of a concept can fall under the concept's extension.

Turning back to the HOT theory of consciousness, I believe that it provides important ammunition against the charge that extant representational theories of mental content fail to account specifically for *conscious representation*, or what has been called "personal level representation" (Georgalis 2006; Kriegel, forthcoming). The complaint is that personal-level representation is a three-place relation (x represents y to S) as opposed to the two-place relation (x represents y) that dominates the literature. And it may well be true that representational theories of content *by themselves* cannot handle, or even ignore, personal-level representation. According to these theories, the process of content acquisition does indeed seem to occur at the unconscious or subpersonal level. But this should not be a surprise, especially if one is inclined to favor a reductionist approach. In my view, this is all the more reason to demand that a *further* metarepresentational level is needed for *conscious* states, which would include a personal-level representation and creature consciousness. This is exactly what HOT theory requires. If we have a plausible causal theory of mental content, but only for unconscious first-order states, then we can see why a HOT is also needed for explaining conscious states and content.[13]

(2) Recall the earlier distinction between Russellian and Fregean contents. Unlike most reductive representationalists, I propose that we should make room for both kinds of content in characterizing a conscious mental state. I see little reason to adopt one at the expense of the other. The contents of conscious states include both Russellian and Fregean elements. Representationalists typically have in mind Russellian contents, but they are not normally thinking in terms of the HOT theory. An advantage of HOT theory is that it can explain how first-order conscious states can embody both kinds of content while retaining its reductionistic credentials.

So the content of, say, a first-order conscious perception is Russellian, but with the help of the relevant HOT, it is also Fregean. We might thus call the content of the resulting complex conscious state *Fregellian*. That is, the HOT will typically tell us *the way* that the objects (or properties) referenced in first-order states *are presented to the subject*. We might say that the *mode of presentation* is normally determined by the HOT's content, that is, the way that the lower-order state is experienced by the subject. Thus, according to Fregellian content, a conscious state can be teased apart in a way that accommodates both Fregean and Russellian content. I qualify and further address the exact nature of the relationship between a HOT and its target in later chapters. Nonetheless this view is still reductive because what accounts for the Fregean content in a conscious state is itself unconscious. This move is not available to first-order representationalists because there is only one level of mental content.[14]

(3) Given the foregoing construal of Fregellian content, it is also natural to allow for *narrow content*, at the least, in addition to wide referential content. Thus I suggest that we should opt for "moderate internalism" (or a "two-factor" theory) as opposed to what is called "extreme internalism" (Segal 2000). The extreme internalist holds that there is only narrow content, whereas the moderate internalist allows for both wide and narrow content. Recall from section 2.1 that we can understand narrow content in terms of whatever it is that molecular duplicates share from the first-person point of view, even if the relevant mental states are also individuated widely. Many who favor narrow content recognize that both narrow and wide contents are legitimate, depending on the context. While it is true that most reductive representationalists are extreme externalists who reject the viability of all narrow content, I believe that this is a mistake. Although it is not always easy to specify the nature of narrow content for concepts and intentional contents, there are compelling reasons to allow for it.[15]

I will not survey all the arguments for and against narrow content (see Brown 2008). My primary focus is on consciousness, not theories of content. Let me briefly offer two reasons to favor narrow content:

(a) Many of us believe that in Putnam's Twin Earth scenario there is still *something* mental that is shared between me and my twin with respect to water thoughts, although our intentional contents might differ when individuated widely. Similarly, suppose that two individuals (P and Q) are having subjectively indistinguishable experiences of an angry tiger, though Q is having a hallucination. One way to capture what they have in common is to resort to narrow content. Indeed, their brains are presumably in very similar states, despite the external differences.

(b) This last point highlights another motivation for narrow content, namely, that it is needed for causal and psychological explanation. For example, P and Q might behave in very similar ways, such as running screaming to safety. Narrowly individuated contents can parsimoniously explain the behavior of both P and Q. Indeed, it is presumably the narrow contents that cause the behavior, though there is not a tiger at all in the case of Q. As Carruthers puts it: "There is every reason to think that psychological laws (or nomic tendencies) should be framed in terms of contents which are individuated narrowly" (2000, 107). Although wide content has its purposes, narrow content is also needed for psychological explanation. It is important to recognize that narrow content can still be accommodated within a reductionist program although many of its proponents in fact reject reductionism.[16]

In conclusion, then, we have made significant progress in establishing the HOT Thesis. Reductive representationalism is a viable strategy to explain consciousness, and HOT theory is a plausible candidate for the task. Intentionality and genuine mental content do not automatically entail consciousness. But much more needs to be done to rule out similar theories of consciousness. I now turn to a critique of several close relatives of HOT theory.

3 Assessing Three Close Rivals

It would be impossible to attempt to refute all, or even most, philosophical theories of consciousness on offer at the present time. The argument of the previous chapter provides the reader with an overall sense of why I reject a number of other theories, such as any nonrepresentational or nonreductionist theory. In this chapter, however, I wish to argue against three theories that are much closer to my own. All of them share the common desire to offer a reductive theory of consciousness in mentalistic terms. In section 3.1, I critically examine first-order representationalism (FOR), including Tye's version of FOR. In section 3.2, I reject Carruthers's dual-content, or dispositional HOT, theory. In section 3.3, I criticize Lycan's higher-order perception (HOP) theory. The reader can think of this chapter as an additional argument by elimination and thus as further support for the HOT Thesis.

3.1 First-Order Representationalism (FOR)

As we saw in chapter 2, FOR refers to theories that attempt to explain conscious experience primarily in terms of first-order intentional states. The two most-cited FOR theories are those of Dretske (1995) and Tye (1995, 2000).

3.1.1 Tye's PANIC Theory

Tye's theory is the most fully worked out FOR theory and so will be the focus of this section. Like others, Tye holds that the representational content of a conscious experience (that is, what the experience is about) is identical with the phenomenal properties of experience. As he put it: "Phenomenal character (or what it is like) is one and the same as a certain sort of intentional content" (Tye 1995, 137). As we saw in the previous chapter, Tye uses the "transparency of experience" to support his view. When one turns one's

attention away from the blue sky and onto one's experience itself, one is still only aware of the blueness of the sky. The experience itself is not blue. One "sees through" the experience to its representational properties, and there is nothing else to one's experience over and above such properties. I do not wish to challenge this argument at this point.[1]

A common initial objection is that FOR does not account for all kinds of conscious states. Aren't there some nonrepresentational conscious states? Some conscious states seem not to be "about" anything, such as bodily sensations (pains, orgasms), moods, emotions, or afterimages. If so, then conscious experience cannot generally be explained in terms of representational properties (Block 1996). Tye responds that pains, itches, and the like do represent in the sense that they represent parts of the body. Either afterimages and hallucinations misrepresent, which is still a kind of representation, or the conscious subject takes them to have certain representational properties from the first-person point of view. He explains that "the qualities of which we are all directly aware in introspecting pain experiences are not qualities of the experiences (assuming that no massive error occurs), but qualities of bodily disturbances in regions where the pains are felt to be" (Tye 2000, 50). The same goes for the other sensory modalities.

Tye also says that felt emotions "are frequently localized in particular parts of the body. . . . For example, if one feels sudden jealousy, one is likely to feel one's stomach sink . . . [or] one's blood pressure increase" (51). He believes that something similar is true for fear or anger. Moods, however, are quite different and not usually localizable in the same way. But if one feels, say, elated, then one experiences an overall change in oneself. Indeed, Tye admirably goes to great lengths to respond to a whole host of alleged counterexamples to FOR. Although I have doubts about some of these replies, I am willing to concede that Tye can at least adequately respond to this kind of objection. Indeed, as a strong representationalist, I am often sympathetic with Tye's overall approach. However, we do not yet have an explanation for what makes a mental state conscious.

Whatever the merits of the argument from transparency and the scope of representationalism (Kind 2003, 2007), it is clear that not all mental representations are conscious. So the key question again becomes: what exactly distinguishes conscious from unconscious mental states (or representations)? What makes a mental state a conscious mental state? Here Tye defends what he calls PANIC theory. The acronym "PANIC" stands for poised, abstract, nonconceptual, intentional content. Tye holds that at least some of the representational content in question is nonconceptual (N), which is to say that the subject can lack the concept for the represented

properties, such as an experience of a certain shade of red that one has never seen before.[2]

Conscious states must clearly also have intentional content (IC) for any representationalist. Tye asserts that such content is abstract (A) and not necessarily about particular concrete objects. This condition is needed to handle cases of hallucinations where there are no concrete objects at all or cases where different objects look phenomenally alike. Perhaps most important for mental states to be conscious, however, is that such content must be poised (P), which is an importantly functional notion. The "key idea is that experiences and feelings . . . stand ready and available to make a direct impact on beliefs and/or desires. For example . . . feeling hungry . . . has an immediate cognitive effect, namely, the desire to eat. . . . States with nonconceptual content that are not so poised lack phenomenal character [because] . . . they arise too early, as it were, in the information processing" (Tye 2000, 62).

This is perhaps where the most serious objection appears. The problem is that what really seems to be doing most of the work on Tye's PANIC account is the extremely functional-sounding "poised" notion, and so he is not really explaining phenomenal consciousness in entirely representational terms (Kriegel 2002a). It is also unclear just how a disposition can confer *actual* consciousness on an otherwise unconscious mental state. Carruthers similarly asks: "How can the mere fact that an [unconscious state] is now in a position to have an impact upon the . . . decision-making process [or beliefs and desires] confer on it the subjective properties of feel and 'what-it-is-likeness' distinctive of phenomenal consciousness?" (2000, 170). As far as I can see, it cannot. Carruthers follows up on his critique of FOR by arguing that it cannot properly distinguish between unconscious and conscious mental states (or between what he calls "worldly subjectivity" and "experiential subjectivity"). He explains that "any first-order perceptual state will be, in a sense, subjective. That is, it will present a subjective take on the organism's environment. . . . But phenomenal consciousness surely involves a much richer form of subjectivity than this . . . and [has] a distinctive *feel* or phenomenology" (2005, 70).

Carruthers offers several elaborate arguments against the plausibility of FOR; it is impossible to discuss them all here. One such argument emphasizes just why FOR theories are not explanatory: it remains mysterious just how the intentional contents in question can be transformed from states with mere worldly subjectivity to states that are phenomenally conscious. In Tye's view, part of what is supposed to do the work is that contents of phenomenal states are *available* to make an impact on one's beliefs and

desires. As we have seen, this is a functional notion, but Carruthers urges, "It just isn't clear why this sort of availability should give rise to the subjective *feel* that is distinctive of phenomenally conscious states" (2005, 103). Carruthers often relies on the two-systems theory of vision (Milner and Goodale 1995) whereby the perceptual states produced by the (human) ventral (or temporal) system are phenomenally conscious, whereas those produced by the dorsal (or parietal) "how-to" action system are not (Carruthers 2005, 72–73, 98–99, 199–201). Although a FO theorist could also accept this theory of vision, it would remain unclear just *why* the ventral states are conscious, since they also produce intentional contents that represent distal properties of the environment and could thus at best produce worldly subjectivity. In some ways, then, Tye's theory is not really a representational theory in the end.

A final related point: It is one thing to say that mental states are conscious *when* they are available to have a direct impact on one's beliefs and desires, but it is quite another to say that such availability is what *explains* state consciousness. The proper explanatory order seems to be reversed. Some mental states are available to have such direct impact *because* they are conscious, *not vice versa*. That is, when I am in a conscious state, such as feeling hungry, it is true that my hunger will directly impact my beliefs and desires. But this is not an explanation of what makes the mental state conscious in the first place. The disposition in question is the *result* of the state being conscious, not the reason why the state is conscious.

It is interesting to note that even some staunch *defenders* of FOR (Byrne 2001, 233–234) concede that it at best offers a *start* on an adequate theory of consciousness. FOR is therefore limited and "isn't much of a theory of consciousness . . . [but] a point from which theorizing about consciousness should start" (234). I suggest that HOT theory is a much more satisfactory theory with the support of the Transitivity Principle and its ability to use HOTs and their concepts to differentiate unconscious from conscious states.

3.1.2 Other Problems for FOR

The deeper issue is that it is not easy to see how *any* FOR can avoid the problem of explaining what differentiates conscious states from unconscious states. It would seem that any FOR "must claim that the difference between [conscious] and [unconscious] mental states is a difference between *what those states represent*" (Kriegel 2002a, 57). But it is difficult to understand how any differences in environmental objects or properties could also mark the difference between conscious and unconscious states. Any environmental feature could be represented unconsciously as well as consciously. This

is precisely why Tye finds it necessary, in the end, to try to explain the difference in some other way, that is, via the dispositional or functional property of "poise." It is interesting to note that HOT theory does not suffer from the same problem. For one thing, according to HOT theory, there are *two* representational layers on which one can construct a theory of state consciousness and thus no reason to appeal to nonrepresentational properties.

There are many other objections to FOR.[3] Historically among them are hypothetical cases of inverted qualia, the mere possibility of which is sometimes taken as devastating to (wide) representationalism (Shoemaker 1982). These are scenarios where behaviorally indistinguishable individuals have inverted color perceptions of objects. Person A visually experiences a lemon in the way that person B experiences a ripe tomato with respect to their colors. The same goes for all yellow and red objects. Isn't it possible that two individuals could exist whose color experiences are inverted in such a way? If so, it would then seem that we have a case where A's visual experiences differ from B's, while the represented objects are the same color.

Perhaps even more relevant here is the Inverted Earth case famously put forth by Block (1990), which is meant to follow up on spectrum inversion. On Inverted Earth, every object has the complementary color that it has here, but we are asked to imagine that a person is unknowingly equipped with color-inverting lenses and then sent to Inverted Earth. Since the color inversions cancel out, the phenomenal experiences remain the same, yet there certainly seem to be different representational properties of the objects involved. Thus the strategy of critics is to put forth counterexamples (either actual or hypothetical) where a difference exists between the phenomenal properties in experience and the relevant representational properties in the world. These objections can perhaps be answered by Tye and others in various ways, but significant debate continues (MacPherson 2005). Intuitions also dramatically differ as to the plausibility and value of such thought experiments.

Although a wide representationalist (who would typically eschew Fregean content) may respond in many ways to Block's scenario (Lycan 1996, 2001b; Tye 2000), it again seems that the most natural response to these cases is simply to acknowledge that there is an unchanging narrow (phenomenal) content when one is transported to Inverted Earth. Those who believe in narrow content can easily handle inverted-spectrum arguments. We can explain the continuity of the Earthling's experience (that is, the unchanging qualia) by appealing to narrow contents. So although wide contents *may* become inverted, the narrow contents remain the same

(Carruthers 2000, 82–86, 109–111). Indeed, inverted qualia arguments are often straightforwardly taken as arguments for narrow content and thus against wide representationalism. Thus I reject so-called phenomenal externalism according to which qualia, or the content of qualitative states, are considered to be wide (Lycan 2001b). Indeed, I remain puzzled by the notion that one's very qualia can be *constituted by*, not merely *caused by*, factors outside the head. Note that this is different from the claim that those mental states *acquire* their content in a way consistent with the causal theory described in the previous chapter.[4]

Another reason to opt for narrow content is based on the idea that there is no such thing as a representation without a mode of representation. We might say that all representation is representation-*as*. As we saw at the end of chapter 2, the mode of presentation can be narrow even when the first-order representational content is itself wide. Let me elaborate again in terms of HOT theory. The *way* that one experiences the contents of one's first-order states is at least partly determined by the concepts in the appropriate HOT. This is also where Fregean content enters into the theory, that is, by distinguishing between sense and reference. Recall also the importance of concepts in shaping one's experiences, such as the wine-tasting example. I said something like this in Gennaro 1996: "One *must* conceptualize one's own conscious states. A conscious state must be presented to its owner in some way or other, i.e., thought of under some mode of presentation" (70). We can still preserve a reductionist program that runs counter to many friends of narrow and Fregean content. Once again, a narrow representationalist could allow for wide content in some cases, such as in Putnam's H_2O and XYZ water thoughts. After all, in those cases, I and my Twin Earth counterpart have indistinguishable conscious experiences *by hypothesis*.

HOT theory has an advantage over FOR in another area: in addition to distinguishing between the vehicle and content of a mental state, we should also note its *modality* (or "attitude") for example, in distinguishing between an auditory and a visual state. It seems possible to have the same content for two mental states with different modalities, such as *seeing* something flying overhead and *hearing* something flying overhead (Kind 2008). But surely these two conscious states have very different qualia although their representational contents are arguably the same: there's something flying overhead. The most obvious way to distinguish these states is through a mode of representation that might seem to bring in a nonrepresentational element. However, this need not be the case. One could instead bring in HOTs to do the job. For example, in the first case, one has a HOT that "I am

seeing something fly overhead" but, in the other, a HOT that "I am hearing something fly overhead."[5]

I now turn to Carruthers's theory.

3.2 Dual-Content Theory

Peter Carruthers (2000) had previously called his theory of consciousness the "dispositional HOT theory" but now refers to it as "dual-content theory" (2005). He believes that it is better to think of HOTs as *dispositional* states instead of the standard view that they are *actual* states. There "need not *actually* be *any* HOT occurring, in order for a given perceptual state to count as phenomenally conscious. . . . The HOTs which render [mental states] conscious are not necessarily actual, but potential" (2000, 227). The basic idea is that the conscious status of an experience is due to its *availability* to higher-order thought. He explains that "phenomenal consciousness consists in a certain sort of intentional content ('analog' or fine-grained) that is held in a special-purpose functionally individuated memory store in such a way as to be available to a faculty of higher-order thought" (2005, 8). Intentional contents are *analog* when they have a finer grain than any concepts that the subject can possess and recall. Thus some first-order perceptual contents are available to a higher-order "theory of mind mechanism," which transforms those representational contents into conscious contents (though no actual HOT occurs).

According to Carruthers, some perceptual states acquire a dual intentional content; for example, a conscious experience of red not only has the first-order content "red" but also has the higher-order content "seems red" or "experience of red."[6] Carruthers often uses consumer, or inferential role, semantics to fill out his theory of phenomenal consciousness. The basic idea is that the content of a mental state depends, in part, on the powers of the organism that "consume" that state, for example, the kinds of inferences that the organism can make when it is in that state. He says that it is because dual-content theory "proposes a set of higher-order analog—or 'experiential'—states, which represent the existence and content of our first-order perceptual states, that the theory also deserves the title of 'higher-order *perception*' theory, despite the absence of any postulated *organs* of higher-order perception" (2005, 64). Thus Carruthers somewhat surprisingly understands his "dispositional HOT theory" to be a form of HOP theory (2004; 2005, chap. 5).

I have argued elsewhere against Carruthers's theory and in favor of a position closer to actualist HOT theory (Gennaro 2004a, 2006b).[7] For my

purposes here, however, I will focus on a number of standard lines of criticism.

First, it remains unclear just what the motivation is to opt for dual-content theory over actualist HOT theory. Carruthers's main objection to the actualist HOT theory is based on what he calls "cognitive overload" (2000, 221–222; 2005, 54). The objection is that actual HOTs would take up too much "cognitive space" (that is, "neural space") given the immense amount that we can experience consciously at one time. Carruthers rejects the reply that our conscious experience is not as rich and complex as it might seem, and thus he believes that dual-content theory fares better on this point. He thinks that since the HOTs are not actual, less cognitive space is needed. However, it is doubtful that this is really the case, regardless of how complex conscious states are. As Carruthers makes clear, dual-content theory still posits the presence of other *actual* structures to fill out his theory, such as a theory-of-mind mechanism and a special memory store. More generally, dispositional states require similar brain structure because something categorical (or actual) must underlie any disposition. No neurophysiological evidence is offered to show that our brains aren't "big enough" to handle the job. After all, it is not just the number of neurons in our brains but also the numerous connections between them.

Second, Carruthers argues that he finds no evolutionary reason to suppose that actual HOTs are present in the case of conscious mental states: "What would have been the evolutionary pressure leading us to generate, routinely, a vast array of [actual] HOTs concerning the contents of our conscious experience?" (2000, 225). But I suggest that Carruthers has overlooked at least three good reasons: (1) according to actualist HOT theory, unconscious HOTs, we may suppose, can more quickly become conscious HOTs resulting in *introspective* conscious mental states. Recall that introspection occurs when an unconscious HOT becomes conscious and is thus directed internally at another mental state. The ability for an organism to shift quickly between outer-directed and inner-directed conscious states is, I believe, a crucial practical and adaptive factor in the evolution of species. For example, an animal that is quickly able to shift back and forth between perceiving other animals (say, for potential food or danger) and introspecting its own mental states (say, a desire to eat or a fear for one's life) would be capable of a kind of practical intelligence that would be lacking otherwise. (2) Even if we suppose that some animals are only capable of first-order conscious states (and thus only unconscious HOTs), the evolutionary foundation has been laid for the yet more sophisticated introspective capacities enjoyed by those of us at the higher end of the evolutionary chain. Thus

the presence of actual unconscious HOTs can be understood, from an evolutionary perspective, as a key stepping stone to introspective consciousness. Such an evolutionary history is presumably mirrored in the layered development of the cortex. (3) Finally, as Rolls points out, actual HOTs allow for the correction of plans that result from first-order processing. Rolls puts forth his own modified version of HOT theory (Rolls 2004) and suggests that part of the evolutionary significance of higher-order thoughts is that they enable correction of errors made in first-order linguistic or in nonlinguistic processing.

Third, a crucial distinction seems lost, or at least unaccounted for, in Carruthers's theory. This is the distinction between first-order (or world-directed) conscious mental states and introspective (or inner-directed) conscious states. Recall that in standard HOT theory, first-order conscious mental states will be accompanied by unconscious HOTs, whereas introspective conscious states are accompanied by conscious HOTs further accompanied by yet higher (third-order) unconscious HOTs. This distinction is noticeably absent from Carruthers's theory, except for one extremely brief mention of third-order states (2000, 251–252). Carruthers needs to answer the following questions: How does he explain the difference between first-order conscious and introspective conscious states on his HOT model? Are the HOTs potential (or dispositional) HOTs only in the first-order case? If so, do they become actual conscious HOTs in the introspective case? If not, how can he account for the difference? Would there be an additional level of dispositional HOTs in the introspective case? Whether Carruthers can answer these questions in a satisfying way without making major modifications to, or even abandoning, his theory is doubtful. Carruthers has told me (via e-mail correspondence) that he is an "actualist" about introspection. Perhaps this is his best option, though it may sound surprising at first. Thus I still wonder (a) why he says so little about the structure of introspection in the context of his theory of consciousness, and more importantly, (b) whether or not this detracts from his initial motivation for a dispositional account of first-order consciousness, especially when combined with the questions raised earlier about his cognitive-overload argument.

Fourth, recall that a key motivation for HOT theory is the Transitivity Principle (TP). But the TP clearly lends itself to an *actualist* HOT theory interpretation, namely, that we *are* aware of our conscious states and not aware of our unconscious states. And, as Rosenthal puts it, "Being disposed to have a thought about something doesn't make one conscious of that thing, but only potentially conscious of it" (2004, 28). Thus it is natural to wonder just how dual-content theory *explains* phenomenal consciousness.

For one thing, it is difficult to understand how a *dispositional* HOT can render, say, a perceptual state *actually* conscious.

To be sure, Carruthers is well aware of this objection and attempts to address it (e.g., Carruthers 2005, 55–60), but his arguments are not convincing. He leans heavily on consumer semantics in an attempt to show that "changes in consumer systems can transform perceptual contents" (56). But the central and most serious problem remains: that is, dual-content theory is vulnerable to the same objection raised by Carruthers against FOR (see the previous section). In both accounts, it is difficult to understand how the functional or dispositional aspects of the respective theories can yield actual conscious states. This point is made most forcefully by Jehle and Kriegel (2006). They rightly point out that dual-content theory "falls prey to the same problem that bedevils FOR: It attempts to account for the difference between conscious and [un]conscious . . . mental states purely in terms of the functional roles of those states" (468). It does indeed seem that if we accept Carruthers's argument against FOR (as I think we should), then it also undermines his own dual-content theory. After all, it is clear that Carruthers intends functional-role semantics to play an essential role in his theory, such that a mental state's content is determined by its functional role in a person's mental life.[8]

Fifth, one might therefore doubt that Carruthers's theory is a theory of *consciousness* at all, as opposed to a theory of *content*, especially since he relies so heavily on consumer semantics to fill out his theory. As Carruthers himself says, "According to all forms of consumer semantics (including teleosemantics and various forms of functional and inferential role semantics) the intentional *content* of a state depends, at least in part, on what the down-stream consumer systems that can make use of that state are disposed to do with it" (2005, 56; italics mine). That is, perhaps consumer semantics can explain why a state has the content it has, but that is not the same as explaining why the state is conscious in the first place. The causal theory of content presented in the previous chapter was intended not to explain state consciousness but to explain how mental states acquire content.

Nonetheless one interesting aspect of Carruthers's view is that conscious states, in some sense, represent themselves. The notion that self-representation or self-reference is an essential aspect of conscious states has a long history, though there are many different versions of this view, ranging from phenomenological (nonreductive) accounts to more recent naturalistic theories. As I remarked in the previous chapter, I am also somewhat sympathetic to this approach going back to Gennaro 1996. That Carruthers sees dual-content theory as a kind of self-referential theory is clear,

for example, when he says that "phenomenally conscious [experiences] . . . come to present themselves to us, as well as presenting properties of the world (or of the body) represented" (2005, 107; cf. 65–66 and chap. 8). We can thus think of a conscious state, in Carruthers's account, as having two contents, one first-order and the other a higher-order reflexive content. If tenable, then Carruthers's account could also be used as a reply to an important objection to actualist HOT theory, namely, that it cannot account for what happens when (or if) the HO state *misrepresents* the LO state (Neander 1998; Levine 2001). As Carruthers says, "It should be obvious why there can be no question of our higher-order analog contents getting out of line with their first-order counterparts, on this account. . . . This is because the higher-order experience *seems [red]* is parasitic on the content of the first-order experience *[red]*" (96). Thus if this version of HO theory guarantees a match between the HO and LO content, then this powerful objection is defused. I discuss this problem at length in the next chapter.[9] Nonetheless I reject dual-content theory for the other reasons offered here.

3.3 Higher-Order Perception (HOP) Theory

David Armstrong (1981) and William Lycan (1996, 2004) are the leading HOP theorists today. I have previously argued that the difference between HOP and HOT theory is greatly exaggerated, though there are perhaps some important differences between thought and perception (Gennaro 1996, 95–101). Van Gulick (2000) has also done us the favor of listing twelve paradigmatic features of perception, including the possibility of error and the lack of personal-level inference between object and perception. But he ultimately concludes that "we are left with no clear judgment in favor of either side" (293). For my own part, I have claimed that what ultimately justifies treating the higher-order states as thoughts is the exercise and application of concepts to lower-order states (Gennaro 1996, 101). As we saw in the last chapter, Rosenthal also relies on this aspect of HOT theory, such as in his wine-tasting example. We also saw in section 3.1 that HOT theory allows for important flexibility not open to FOR (or to HOP theory, for that matter). I develop this theme further in later chapters, for example, by exploring the problem of concept acquisition, by solving the hard problem, and by endorsing conceptualism. Thus the full advantages of HOT theory over HOP theory will be fully apparent only after reading the entire book.

One standard objection to HOP theory is that, unlike outer perception, no obvious distinct sense organ or scanning mechanism is responsible for HORs. Similarly, no distinctive sensory quality or phenomenology is

involved in having HORs, whereas outer perception always involves some sensory quality.[10] Lycan concedes the disanalogy but argues that it does not outweigh other considerations favoring HOP theory (Lycan 1996, 28–29; 2004, 100). Lycan's reply might be understandable, but the objection remains an obvious and crushing one nonetheless. I do not think that this disanalogy can be overstated. After all, this represents a major difference between normal outer perception and any alleged inner perception, which arguably involves the most central characteristic of perception.

Let me also respond briefly to two points raised by Lycan (2004), keeping in mind the distinction between unconscious HOTs (or HOPs) and conscious HOTs (or HOPs).

(1) I do not quite understand why Lycan thinks that there is an "intuitive priority" for HOP over HOT theory (2004, 101). For one thing, as we saw in the last chapter, the Transitivity Principle and Lycan's (2001a) argument are designed to show that *some* HOR theory is intuitively true. If there is any advantage gained from Lycan (2001a), I think that it belongs to HOT theory, since he emphasizes the "of" in "aware of" as the "of" of intentionality. While it is true that perceptual states are also intentional states, thoughts are surely more paradigmatic examples of intentional states.

Lycan (2004, 101) insists, however, that it is "hard to imagine . . . S thinking about X and *thereby* becoming aware of X," whereas it makes perfect sense to say that S's (perceptual) awareness of X must precede any thoughts about X. I am puzzled by this. In five seconds, I am going to think about my last perception of the Empire State Building, and then in ten seconds I am going to think about my desire to see it again. Okay, five, four, three, two, one: I am now surely thereby aware of my last perception of the Empire State Building . . .And now, five seconds later, I am aware of my desire to see it again. Surely my thoughts *made me* aware of those states and preceded my awareness of them. Of course, we are here presumably talking about conscious HOTs, but the point remains, since the issue is which HO theory is closer to the truth. Thinking about X can clearly make one aware of X. I don't see the problem. The same could even hold for outer perception: I am now thinking about my printer, so then I turn my attention to it by looking at it. The thinking of X caused me to become aware of X. It is obviously true that that my thinking about outer objects does not bring them into existence, but then the same is true for perceiving outer objects.

Indeed, to go even further, consciously thinking about (or attending to), say, a pain in one's foot can even make the pain *worse* or bring it into greater prominence (Hill 1991; Gennaro 1996, 20–21). This is what Hill calls "volume adjustment," but he also recognizes what he calls "activation." Activation "occurs if one succeeds in actualizing or activating a sensation

of the right sort" (Hill 1991, 121). It seems to me that these are precisely cases where consciously thinking about X precedes X and brings X into existence. This is consistent with, and further evidence for, the notion that concepts in HOTs can alter the nature of, or even bring into existence, one's conscious states. Lycan agrees that this is an advantage of HOT theory over HOP (2004, 108).

(2) This line of reasoning is also importantly related to Lycan's claim that HOP theory is superior because, by analogy to outer perception, there is an important nonvoluntary or passive aspect to perception not found in thought. But, again, this is precisely the *problem* for HOP theory. The perceptions in HOPs are *too* passive to account for Hill's volume adjustment or activation. Thus HOTs are preferable. The higher-order awareness in question is clearly not analogous to a purely passive outer perception. Even at the level of first-order conscious states (and thus unconscious HOTs), HOT theory can explain the important interaction between the HOT and the first-order state.

I have previously put a similar point in Kantian terms (Gennaro 1996, 45–48): we might distinguish between the faculties of *sensibility* and *understanding*, which must work together to make experience possible. What is most relevant here is that the passive nature of the sensibility (through which outer objects are given to us) is contrasted with the active and more cognitive nature of the understanding, which thinks about and applies concepts to that which enters via the sensibility. HOTs fit this latter description well. In addition, Kant uses the term *Begriff* for "concept." *Begriff*, unlike its English counterpart, is a more active term meaning "a grasping." It has as one of its cognates *begreifen*, which is the verb meaning "to grasp," thus emphasizing the active role played by the understanding. Once again, the wine tasting and similar examples seem to be instances of the way in which one's HOTs (with their constitutive concepts) can affect one's first-order experiences.[11]

(3) Despite the passivity of HOPs, Lycan (2004, 102–105) thinks that HOP theory can better handle the fact that we can voluntarily control HORs, at least when we introspect our mental states. Just as we can voluntarily direct our attention to outer objects through perception, so we do the same to our mental states via HOPs.

In response, as Lycan (2004) recognizes, I had posed the following dilemma for HOP theory: The HOR in question must either be conscious or unconscious. (a) If it is conscious (and thus a case of introspection), then HOTs can at least equally be used to voluntarily and actively search for their mental objects. (b) If it is unconscious, then there is no sense in which one voluntarily controls the HORs in question. So HOP theory cannot have an advantage over HOT theory on this point.

Lycan's reply to (a) is, first, not to concede that the HOT theorist can equally talk of "actively searching" for one's own mental states. For example, he says that we can at will "selectively attend to environmental region R and see whatever there is in R. We do not *in the same facile way* control what things in the environment we have thoughts about; thought is more spontaneous and random" (2004, 104). Lycan continues: "The same goes for the voluntary control of attending to the first-order mental states. At will, we can selectively attend to phenomenal region R and detect whatever sensory qualia there are in R. We do not in the same facile way control what regions of or things in the phenomenal field we have thoughts about" (105).

My reply here is twofold: First, as we saw with my earlier printer example, I can indeed voluntarily control that to which I attend via conscious thinking. It is true that I cannot control what the *object* of my perception will be like once I turn my attention to the printer. But, again, I do not see why Lycan thinks that attention to X must precede thoughts about X. I can think that I will attend to whatever is in region R in my visual field in five seconds, and then do so. Second, and more to the point, if the HORs are conscious, I can equally consciously think about any of my mental states at will. As in the earlier Empire State Building example, I can actively search for my last perception (or memory) of it, as well as my desire to see it again. The same goes for the deliberative introspection of my current beliefs and intentions. As a matter of fact, this is precisely the sort of voluntary searching activity associated with reasoning. It may be that thoughts are generally more spontaneous and random than perceptions, but it does not follow that *when one is voluntarily having a conscious HOR*, one is more likely having a HOP than a HOT. Perhaps we simply disagree about the extent to which we can actively search for one's mental states via thought.

Lycan's reply to (b) is first to concede that there is no voluntary control or attention for unconscious HORs, and thus his argument "does not show that the representation produced is *in its own nature and structure* more perception-like" (2004, 105). So far, so good. But Lycan then still insists that since the "relevant higher-order representations are characteristically produced by the exercise of attention . . . that makes them more like perceptions than like thoughts" (105).

This response also has problems. First, the relevant HORs in question—namely, unconscious HORs—are precisely those *not* produced by the exercise of attention, unless one means some kind of unconscious attention. Just as such unconscious HORs do not admit of voluntary control, so too they do not involve the exercise of conscious attention. Second, it seems that we now have another dilemma: such HO attention is either conscious

or unconscious. (a) If it is unconscious, then, again, there is neither conscious attention directed at the mental state nor any voluntary control involved. In cases of first-order conscious states, one's attention is directed outward. Moreover, it is not even clear what "unconscious attention" is supposed to mean in this context. (b) If it is conscious, then we are no longer talking about cases of unconscious HORs, which is what this horn of my initial dilemma was supposed to be about. We can then revert to my reply to the other horn of the dilemma.

Another often ignored area that favors HOT theory has to do with the way that we normally characterize introspection. I believe that we should distinguish between two kinds of introspection: *momentary focused* and *deliberate* (Gennaro 1996, 18–21). Momentary focused introspection is merely having a momentary conscious awareness of one's own mental state. Deliberate introspection, however, involves sustained conscious attention directed at one's own mental states, for example, when one is philosophizing about one's beliefs or trying to figure out how to solve a problem. This clearly involves reasoning, making inferences, and planning. Unlike perception, this kind of activity clearly does not merely involve passively becoming aware of one's own mental states. It is an engaged, active, and (yes!) often voluntary process that involves manipulating and sometimes altering one's own beliefs, intentions, and desires. This type of introspection is something that we are all familiar with and has much more in common with thinking (via the use of concepts) than with anything like perception. Moreover, it stands to reason that deliberate introspection is simply an extension of momentary focused introspection, which, in turn, is merely the conscious counterpart to unconscious HOTs. Recall also from the previous chapter that some non-HOT theorists are sympathetic to the way that HOT theory defines introspection in terms of conscious HOTs directed at mental states.

Much more could be said about the nature of introspection, but I will not address the overall nature or epistemic aspects of introspection here. We should certainly agree that there is little reason to accept the view that introspection is infallible. I certainly do not hold to the Cartesian "self-intimating" view whereby if one has a mental state, then one automatically knows that one is in it. This would in essence rule out the possibility of any unconscious mental sates.

In conclusion, then, I think that HOT theory is superior to the other three theories discussed in this chapter. The HOT Thesis is now in an even stronger position than at the end of chapter 2. I now turn to a further defense of HOT theory, including my own intrinsic version, the WIV.

4 From HOT Theory to the Wide Intrinsicality View

Having defended reductive representationalism and HOT theory in the previous two chapters, I now motivate and further defend a modified version of the theory, which I have called the wide intrinsicality view, or WIV (Gennaro 1996, 2006a). In section 4.1, I introduce what appears to be a false dilemma invoked by Rosenthal and offer some initial rationale for favoring the WIV over his version of HOT theory. In sections 4.2 through 4.4, I address what I take to be the three most serious objections to standard HOT theory: the problem of misrepresentation, the problem of the rock, and the hard problem of consciousness. As we will see, these problems are related in important ways. More specifically, in sections 4.2 and 4.3, I show why the WIV can better answer the misrepresentation and rock problems. In section 4.4, I show how either version of HOT theory can solve the so-called hard problem of consciousness (Chalmers 1995). I also argue that they are immune to Chalmers's criticisms of other reductionist accounts of consciousness. Finally, in section 4.5, I reply to numerous objections specifically aimed at the WIV and develop significant details of the theory. Thus, overall, I further defend the HOT Thesis as well as the Hard Thesis.

4.1 A False Dilemma

According to the WIV, what makes mental states conscious is *intrinsic* to conscious states, but a kind of *inner* self-referential and relational element is also present *within* the structure of such states. In contrast to standard HOT theory, the WIV says that *first-order* conscious mental states are *complex* states containing both a world-directed mental state-part M and an unconscious metapsychological thought (MET). It is, if you will, an intrinsic version of HOT theory (see fig. 4.1).

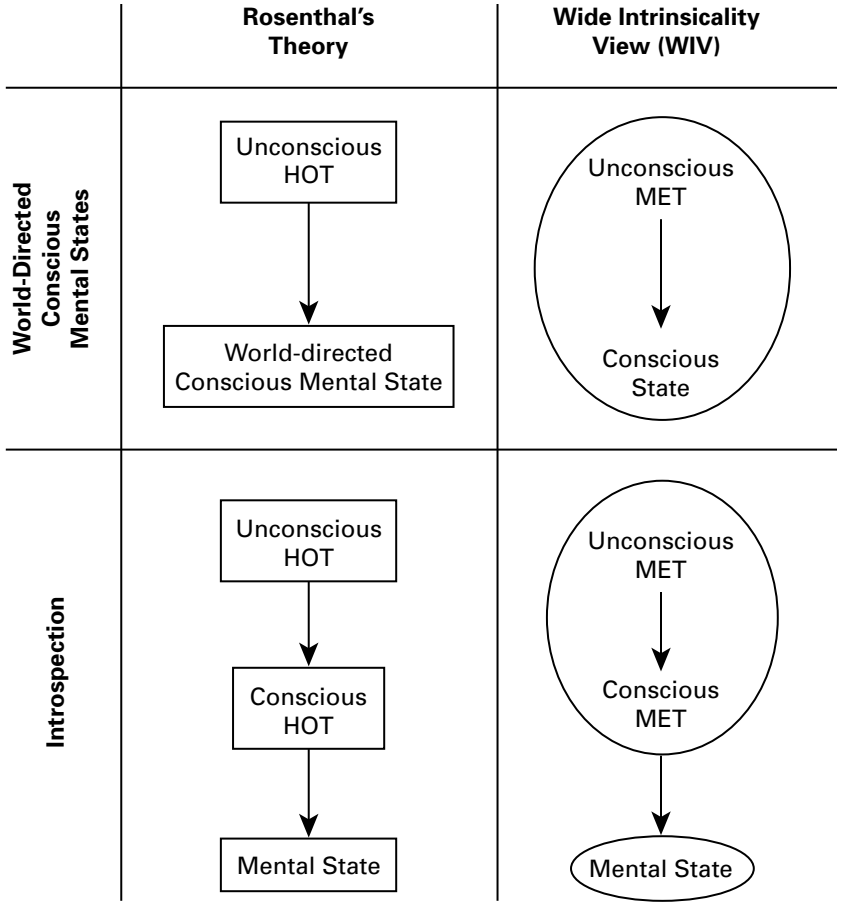


Figure 4.1
The structure of conscious mental states contrasting Rosenthal's HOT theory and my wide intrinsicality view (WIV). "MET" stands for "metapsychological thought."

It is important to mention here just a few reasons to favor the WIV over standard HOT theory. I had originally offered five but will not repeat them all here (Gennaro 1996, 26–30).

First is simply that consciousness seems to be an intrinsic property of conscious states. As Rosenthal has long acknowledged (1986, 331, 345), it is preferable to have a theory that can account for this fact if at all possible. When we are in a conscious state, consciousness does not seem to be analogous to "being the cousin of" or "being to the left of" but instead seems to be part of the state itself. Of course, most would agree that the reality of conscious states need not match the first-person appearance. Rosenthal rightly explains that HOT theory can still accommodate the phenomenological facts by noting that we are rarely conscious of the HOT itself that renders M conscious. Thus Rosenthal says that we should not assume that "consciousness reveals everything about our mental functioning, or at least everything relevant to the issue at hand. . . . But to save these phenomena, we need only explain why things appear to consciousness as they do; we need not also suppose that these appearances are always accurate" (2004, 31).

Fair enough. But the problem is that Rosenthal never really presents a compelling case to reject the intrinsicality of consciousness in the first place. For example, he argues that if we treat consciousness as an intrinsic property of mental states, then conscious mental states will be simple and unanalyzable (see Gennaro 1996, 21–24, for further critical discussion). Rosenthal defines an intrinsic (as opposed to extrinsic) property as follows: "A property is intrinsic if something's having it does not consist, even in part, in that thing's bearing some relation to something else" (Rosenthal 1997, 736). But even if consciousness is an intrinsic property of some mental states, it surely does not follow that those states are simple or unanalyzable. Conscious mental states can, for example, have the kind of complex structure described by the WIV. Rosenthal sets up a false dilemma: *either* accept the Cartesian view that mental states are essentially and intrinsically conscious (and so unanalyzable) *or* accept his version of the HOT theory whereby consciousness, or the so-called conscious-making property (i.e., being the object of a HOT), is an extrinsic property of mental states. But there is clearly an informative third alternative whereby the MET is part of the overall structure of a conscious mental state.

Furthermore, it is not clear that one must rely on any such phenomenological or "intuitive" evidence to make the point. If we examine the issue from a third-person neurophysiological perspective, there is still something odd in holding that what makes a mental state M conscious is *something*

else. For example, if and when the true neural correlate(s) of consciousness are discovered, it seems far more likely that it (or they) will be treated as part of the conscious brain state. That is, what makes a mental state conscious will be some (perhaps distributed) property of the state itself. There can still be a self-referential structure to that conscious state, but both M and MET will be parts of the overall state.

Rosenthal (2004, 33) rightly demands that an "intrinsic theory must explain what happens when a state goes from being nonintrospectively conscious to being introspectively conscious." But, as he knows, I had already presented the WIV version of introspection in both Gennaro 1996 and 2002. Once again, according to the WIV, *first-order* conscious mental states are *complex* states containing both a world-directed mental state-part M and an unconscious metapsychological thought (MET). My conscious perception of the tree is accompanied by a MET *within* the very same complex conscious state. Now when I *introspect* my perception, a first-order mental state is rendered conscious by a complex higher-order state. Thus *introspection* involves two states: a lower-order noncomplex mental state that is the object of a higher-order conscious complex state (see fig. 4.1 again). In this case, much like in HOT theory, the MET itself becomes conscious and is directed at a lower-level mental state.

Another reason to favor the WIV is that intrinsic theory seems better suited to avoid two major problems facing standard HOT theory. (1) Rosenthal's HOT theory arguably has a more serious problem dealing with the possibility of *misrepresentation* between the HOT and its target state M (Byrne 1997; Neander 1998; Levine 2001). If we are dealing with a representational relation between two *distinct* states, it is possible for misrepresentation to occur. In any form of intrinsic theory, such as the WIV, it seems more difficult to make sense of the possibility of misrepresentation, since the MET is directed back at a mental state M, which is part of the same state as M. I will address this problem at length in section 4.2 and again later in section 4.5.

(2) Another well-known difficulty for standard HOT theory has been called "the problem of the rock" (Stubenberg 1998) and "the generality problem" (Van Gulick 2006). When I have a thought about a rock, it does not thereby make the rock conscious. So why should we suppose that when one thinks about a mental state M, it becomes conscious? This problem is again based on the notion that the HOT is completely separate from the target mental state M. If this aspect of HO theory is rejected, then its target must be a mental state "in the head" in the first place, and so objects such as stones cannot be potentially conscious.[1] I return to this issue in section 4.3.

Having said that, it is necessary to digress for a moment to clearly establish the relative logical space that the WIV occupies, especially given the potential for terminological confusion. First, one will notice that the WIV is extremely similar in structure to Rosenthal's HOT theory. I do indeed think that there is something importantly correct about HOT theory, and my view admittedly owes much to Rosenthal's theory. It may even seem that no real ontological difference distinguishes the two views, but this is not correct for reasons that I clarify later in this chapter. Second, *some* of the same objections to standard HOT theory might also be raised against the WIV, so I have responded to them in print. As I mentioned in chapter 1, I have argued at length against the view that HOT theory entails a lack of animal consciousness (Gennaro 1993, 1996, 2004a) and against other weak attempts to criticize HOT theory (Gennaro 2003). Thus, in these cases, I have no problem associating myself with some form of HOT theory and thus willingly bring such a characterization upon myself. Of course, if one *defines* HOT theory as maintaining that the HOT is a *distinct* state from its target M, then it would *not* be correct to say that I hold a HOT theory. I do *not* wish to define HOT theory *as such* in this way, but others clearly do. In my view, this is more of a terminological dispute, though some seem to have strong views on the matter.[2]

On the other hand, some have also urged me to reject HO theory altogether and endorse a first-order theory, perhaps more along the lines of Dretske (1995) or Tye (1995, 2000).[3] However, as I have made clear in the previous chapter, I am not inclined to do so and reject FOR theory. In any case, I take such accusations from both sides as evidence that the WIV is a truly viable "middle" position between FOR and standard HOT theory. Indeed, that is precisely why I originally chose to introduce a new theory name into the literature. It is also why I sometimes use the more neutral expression "metapsychological thought" (MET) instead of "higher-order thought" (HOT), especially in this chapter: there is still a MET about M in the WIV.[4]

4.2 Misrepresentation: A First Pass

4.2.1 Levine's Case

With regard to the problem of misrepresentation, I focus first on the way that Levine (2001) presents this objection against all HO theories. He credits Neander (1998) for an earlier version of this objection under the heading of the "division of phenomenal labor." The idea is that when "we are dealing with a representational relation between two states, the possibility of

misrepresentation looms" (Levine 2001, 108). Levine argues that standard HOT theory cannot explain what would occur when the higher-order (HO) state misrepresents the lower-order (LO) state. The main example used is based on color perception, though the objection could presumably extend to other kinds of conscious states. Levine says:

> Suppose I am looking at my red diskette case, and therefore my visual system is in state R. According to HO, this is not sufficient for my having a conscious experience of red. It's also necessary that I occupy a higher-order state, say HR, which represents my being in state R, and thus constitutes my being aware of having the reddish visual experience. . . . Suppose because of some neural misfiring (or whatever), I go into higher-order state HG, rather than HR. HG is the state whose representation content is that I'm having a greenish experience, what I normally have when in state G. The question is, what is the nature of my conscious experience in this case? My visual system is in state R, the normal response to red, but my higher-order state is HG, the normal response to being in state G, itself the normal response to green. Is my consciousness of the reddish or greenish variety? (Levine 2001, 108)

Levine initially points out that we should reject two possible answers:

Option 1: The resulting conscious experience is of a *greenish* sort.
Option 2: The resulting conscious experience is of a *reddish* sort.

I agree that options one and two are arbitrary and poorly motivated. Option one would make it seem as if "the first-order state plays no genuine role in determining the qualitative character of experience" (Levine 2001, 108). The main problem is that one wonders what the point of having *both* a LO and HO state is if only one of them determines the conscious experience. Moreover, HOT theory is supposed to be a theory of (intransitive) state consciousness; that is, the *lower-order* state is supposed to be the conscious one. On the other hand, if we choose option two, then we have the same problem, except now it becomes unclear what role the HO state plays. It would then seem that HOTs are generally not needed for conscious experience, which would obviously be disastrous for any HO theorist. Either way, then, options one and two seem to undermine the relational aspect of HOT theory. Thus Levine says: "When the higher-order state misrepresents the lower-order state, which content—higher-order or lower-order—determines the actual quality of experience? What this seems to show is that one can't divorce the quality from the awareness of the quality" (2001, 168).

It is important to point out here that Rosenthal defends Levine's option one. For example, with respect to "targetless" HOTs, where there is no LO state at all, Rosenthal explains that the resulting conscious state might just be *subjectively indistinguishable* from one in which both occur (Rosenthal

1997, 744; cf. 2005, 217). I find this view highly implausible, as I have already mentioned. It also seems to me that since the HOT is itself unconscious, there would not be a conscious state at all unless there is also the accompanying LO state. We would merely have an unconscious HOT without a target state, which by itself cannot result in a conscious state. Levine says, "Doesn't this give the game away? . . . Then conscious experience is not in the end a matter of a relation between two (non-conscious) states" (2001, 190). On the other hand, I argue that the self-reference and complexity of conscious states in the WIV rule out this kind of misrepresentation. If we have a MET but no M at all (or vice versa), then what would be the *entire* conscious state does not exist and thus cannot be conscious. A CMS will exist only when its two parts exist and the proper relation holds between them.

Returning to the foregoing example, both Levine (2001, 108–109) and Neander (1998, 429–430) do recognize that other options are open to the HO theorist, but they quickly dismiss them. I focus on Levine's treatment of these alternatives and argue that they are more viable than he thinks.

Option 3: "When this sort of case occurs, there is no consciousness at all" (Levine 2001, 108).
Option 4: "A better option is to ensure correct representation by pinning the content of the higher-order state directly to the first-order state" (Levine 2001, 108).

First, we must be clear about what Levine means, in option three, by "no consciousness at all." Presumably he does not mean that the hypothetical *person* in question would be completely unconscious. This would be a puzzling consequence of any HO theory and would also confuse creature consciousness with state consciousness. So it would seem that Levine's option three is really saying that, in such cases of misrepresentation, the person has *neither* the greenish *nor* the reddish conscious experience. But then it becomes unclear why Levine rejects option three as ad hoc (2001, 108). What exactly is so ad hoc about that reply? HOT theory says that when one has a conscious mental state M, it is accompanied by a HOT that "I am in M." If there isn't a "match" (that is, an accurate representation) between a HOT concept and the content of a lower-order state, then it seems perfectly appropriate for the HOT theorist to hold that something like option three is a legitimate possibility. After all, this is an abnormal case where applying the HOT theory could not be expected to result in a normal conscious state. We are not told just how unlikely or unusual such a scenario is. Levine's thought experiment lacks an important level of detail. Recall

that we are simply told to "suppose because of some neural misfiring (or whatever)." Perhaps there would be no resulting conscious color experience of the diskette case at all. Alternatively, if specific brain lesions are involved, perhaps the subject would at least experience a loss of color vision (achromatopsia) with a diskette case perception. It seems to me that option three is not at all implausible: if misrepresentation occurs between M and MET, then no conscious state results. At the very least, if a misrepresentation occurs between some of the relevant concepts in M and MET, then *that aspect* of the conscious state would not exist. To use an oversimplistic analogy: if I call my friend Tom and there is something wrong with his phone or the connection, then no phone conversation will take place.

This brings us to the so-called better option in option four, which also seems plausible and sounds very much like the WIV. In a sense, defending option three might lead one naturally to option four. Indeed, they seem to be two sides of the same coin because option three is also, in essence, claiming that a match, with respect to the relevant concepts involved, between a HOT (or MET) and a lower-order state must be "ensured" to result in a conscious experience. Levine does mention two problems with this fourth approach, but I am puzzled by his remarks. He *first* asks, "What if the higher-order state is triggered randomly, so that there's no first-order sensory state it's pointing at? Would that entail a sort of free-floating conscious state without a determinate character?" (Levine 2001, 109). I will discuss targetless HOTs later in this section, but the answer to Levine's second question is clearly no, because you would merely have an *unconscious* HOT without a target state, as was noted earlier. An unconscious HOT, by itself, cannot result in a conscious state of any kind unless a yet further third-order state is directed at it. *Second*, Levine simply expresses puzzlement as to how option four "overcomes the basic problem . . . [which] is that it just doesn't work to divide phenomenal labor" (Levine 2001, 109; cf. Levine 2006). But this is not really *another* objection aimed at option four or HOT theory—it merely repeats Levine's conclusion.

In addition, when Levine says that my "visual system is in state R . . . but my higher-order state is HG," this is misleading and perhaps even begs the question against the HO theory. What encompasses the "visual system"? Levine assumes that it is only the lower-order state R. However, if HOT theory (or the WIV) is true, it seems much more plausible to treat the *entire* system (including both the lower-order and higher-order state) as parts of the visual system in this case. Thus the visual system (or at least the *conscious* visual system) would have to contain R *and* HR, so that there *would be* a conscious reddish experience (even if an idle HG state also exists). Perhaps

option two is thus not so arbitrary after all. At the least, the hypothetical scenario would seem to be misdescribed or just assumes the falsity of the HOT theory. HOTs (or METs) should be understood as part of the visual system when one is having a first-order conscious perception of any kind. In this case, then, we should also say that R and HR are each *necessary* for having a reddish experience, but neither one is *sufficient* by itself. R and HR are jointly sufficient.

Nonetheless I think that Levine and Neander have, in a somewhat indirect way, hit upon one important and potentially troubling issue regarding the nature of HOT theory. Levine's argument may indeed contain a grain of truth, namely, that it is difficult to make sense of *entirely* splitting off the lower-order state from the HOT, as standard HOT theory claims. Thus he is grappling with a deeper issue that must be addressed by any HOT theorist. It is perhaps best expressed when Levine says that HOT theory has difficulty with "the paradoxical duality of qualitative experiences: there is an awareness relation, which ought to entail that there are two states serving as the relevant relata, yet experience doesn't seem to admit of this sort of bifurcation. Let's call this the problem of 'duality'" (Levine 2001, 168).

The problem of duality (and misrepresentation) is important, but I think it can ultimately be handled best by adopting the WIV. The WIV can help to alleviate some of the puzzlement expressed by both Neander and Levine. For example, a proponent of the WIV can respond that conscious experience (from the first-person point of view) does not seem to allow for a split ("bifurcation") between the lower-order and higher-order states. However, the "awareness relation" does not entail the existence of two entirely separate states. Instead, according to the WIV, we have two parts of a single conscious state with one part directed at ("aware of") the other. In short, we have a complex conscious mental state with an inner intrinsic relation between parts. There is thus a kind of "self-referential" or "self-representational" element to conscious states. This element of self-reference seems to rule out the kind of misrepresentation that threatens HOT theory, because if a MET is misrepresenting M (or if there is no M at all), then what *would be* the proper *entire* conscious state does not exist and thus cannot be conscious. A CMS cannot represent itself (or part of itself) if it doesn't exist in the first place. According to the WIV, a CMS will exist only when its two parts exist and the proper relation holds between them. Moreover, the MET refers back to a part of the CMS (of which MET is also part), so there would be no complex CMS if there is no M at all or if the MET is somehow inaccurately representing M. In standard HOT theory, no such claims can be made because the M and HOT are entirely distinct existences.

In any case, Neander credits Barry Loewer for an "ingenious suggestion that might be worth pursuing":

> The suggestion is that the two levels of representation might be collapsed into one level that is *self-referential*. . . . This suggestion also rids us of the division of phenomenal labor, while still allowing us to maintain that the difference between conscious and unconscious sensory representations is that some of them are meta-represented and some are not. Since the first and second-order representings no longer involve two separate representations . . . the two cannot come apart, so mis-(meta-)representation is in principle impossible. (Neander 1998, 429–430)

This sounds familiar, but Neander unfortunately also dismisses this option too quickly. As I have argued, I think it is a truly viable option that can help to counter Levine's problem of duality. Consider the claims that, according to the WIV,

(1) There is no resulting conscious state when a misrepresentation does occur, and

(2) Misrepresentations cannot occur.

I am honestly unsure which view is preferable, but either one seems plausible. Indeed, (1) and (2), like Levine's options three and four, seem to be two sides of the same coin. As a practical matter, it doesn't really matter very much. Statement (2) should really be understood as

(2′) Misrepresentations cannot occur *between M and MET and still result in a conscious state*.

It is now possible to address Levine's related concerns about whether or not qualia can be explained as *either* intrinsic *or* relational features of conscious states (2001, 93–107; 1995). With the WIV alternative, we can again clearly see the false dichotomy. Qualia are qualitative properties of qualitative states, which are complex states with *both* intrinsic *and* (inner) relational features. This solution is perhaps similar to what Levine calls "the complexity gambit" (2001, 95), but I have already argued that such a move can address his challenge about the nature of qualitative states while at the same time rejecting the notion "that no progress can be made [in explaining consciousness] if we consider qualitative character to be an intrinsic property of experience" (Levine 2001, 94). Once again, this is an advantage of the WIV over Rosenthal's HOT theory: consciousness can be both an intrinsic and relational property of experience without giving up on explaining its nature. We need not hold that treating consciousness as an intrinsic quality of experience forces one to adhere to the implausible

Cartesian position such that consciousness is an unanalyzable property of conscious states.

Moreover, my variation of HOT theory allows us to avoid a controversial aspect of Rosenthal's theory that is also a target of Levine's critique: Rosenthal's theory "splits off subjectivity from qualitative character, and . . . it is precisely this feature that seems so implausible" (Levine 2001, 105). Unlike Rosenthal (1991, 2005), I do not hold that there can be, for example, unconscious *sensory* or *qualitative* states. Some of the disagreement here is purely terminological, such as how to use the terms "sensory," "experience," and the like (see chapter 1). However, there is also substantial disagreement. Since Rosenthal believes that HOTs are extrinsic and inessential to their target states, he (unlike me) holds that the lower-order states can exist without HOTs *and continue to have their qualitative properties*. According to Rosenthal, an unconscious state can thus have "qualitative character," but not "subjectivity" (or "what it's like for the subject"), which requires a HOT. The HOT merely "reveals" the already existing qualitative property to the subject.

To be fair, Rosenthal develops a detailed account of these qualitative properties (or "sensory qualities"), which he calls "homomorphism theory," whereby unconscious states resemble and differ from one another in ways that reflect the resemblances and differences among perceptible properties, such as color and shape.[5] We might classify these qualities in terms of families of properties that pertain to color, shape, sound, and so on (Austen Clark 2000). For example, "The red sensory quality of visual sensations resembles the orange sensory quality of such sensations more than either resembles the sensory green or blue of each sensation" (Rosenthal 2005, 140). Rosenthal argues that much the same is true for shapes, bodily sensations, and other kinds of sensory qualities. "Mental roundness and triangularity resemble and differ from each other in ways homomorphic to the similarities and differences that hold between physical roundness and triangularity" (140). And since these resemblances and differences are also found in conscious sensations and perceptions, Rosenthal infers that those qualitative properties are *the same as* those that figure into unconscious states. The properties of being conscious and having sensory quality are independent of each other.

But one problem for Rosenthal's homomorphism theory is again apparent in cases of misrepresentation. For example, if a HOT misrepresents its target state M and Rosenthal accepts Levine's option one, then he is committed to the view that the HOT's color "qualitative property" (= green) becomes conscious, not the color quality in M (= red). But this contradicts Rosenthal's clear desire to keep the unchanging qualitative property at the

lower-level M. Further, if we wish to hold that HO conceptualization is *essential* to the identity of a conscious state, then an unconscious qualitative state could not be the *very same state* as its conscious counterpart. In my view, the MET's relation to M must play a role in determining the very qualitative properties of a conscious state. There *must be something present* at the LO level for the MET to *recognize* the LO input. When the MET recognizes a LO input as having the same or similar concepts, the result is a conscious (qualitative) state. If my LO state registers a red percept and it is recognized as such in a MET, then the conscious experience will reflect that fact. The same goes for any object or property concept. So, for example, on my view there can be unconscious perceptions and pains, but they are not "sensory" or "qualitative" states. When a pain or perception (M) is properly recognized by a MET, then the combination of M and MET becomes a conscious state by virtue of M becoming the target of an appropriate MET. The entire complex state is then a qualitative state. For reasons that will become clear, we will need to revisit the misrepresentation problem in section 4.5 and in chapter 6.

4.2.2 An Objection

For now, however, one might object that, by analogy, self-reference does not normally rule out misrepresentation. For example, think about self-reference within a sentence. Consider the following:

(S) This sentence ends with a five-letter word.

S is not only false; it is false because of a kind of self-referential misrepresentation, namely, the final word in the sentence. That is, the word "word" has four letters. Why can't something similar be true for the MET according to the WIV? A MET could misrepresent M even if they are both parts of a larger complex in a way analogous to S. Much the same is true for, say, a painting of someone painting a portrait of a woman, Jane, where the painter *in the painting* is *misrepresenting* the color of Jane's dress. Indeed, it doesn't even help if the self-reference is directed at the *entire* sentence:

(S′) This sentence has three words.

My reply is threefold. First, I am not quite sure what to make of these analogies. After all, if consciousness is special and unique in many ways, we should not expect it to be so easily compared to superficially related objects or linguistic expressions. Analogies are plausible only if the compared items do not have too many dissimilar respects. Conscious states are not very much like sentences, though we can have thoughts with the same content.

So I reject the assumption that sentences or paintings are sufficiently like conscious states to draw any legitimate conclusion about the structure of conscious states. It is also unclear what is alleged to be the analogue of having a CMS in the example sentences. If it is merely the existence of the sentence, then that is one thing and would be unproblematic. However, perhaps the analogue of the entire CMS, if there is one at all, ought to be the *truth of S or S'*. If this is the case, then, like Levine's option three, no such product results even if we allow for misrepresentations within sentences. And what exactly is supposed to be the analogue of the CMS in the painting? It is not clear to me. Here we have images that are even less analogous to conscious states.

Second, in the sentence case, it is unclear what is the analogue of the *psychological integration* between M and MET that I will describe throughout this chapter. No psychological interaction occurs between the parts of S or parts of a painting. Third, sentences are not causally efficacious events realized in complex brain activity, as conscious states presumably are. It is difficult to understand how to compare linguistic expressions or images in a painting to neural events. No causal interaction occurs between the parts of S or between parts of a painting. At best, a more (but not very) analogous situation to the WIV might be the following:

(S″) This sentence ends with a period

Notice that there is no period at the end of S″, and thus there is a misrepresentation in S″ (and so it is false). However, we might also counter that S″ is not a proper sentence in the first place because proper sentences end with periods. If "being a sentence" is meant to be analogous to "being a conscious state," then we might say that *just as* the entire state would not be conscious if a MET misrepresents M, *so* S would not really be a sentence.

Of course, one could then alter S″ to read as the following false sentence:

(S‴) This sentence does not end with a period.

So much for analogies. No doubt other similar sentences could be offered, such as a sentence with two conjuncts, the first of which misrepresents the second. But it again seems to me that the dissimilarities far outweigh the similarities. Thus I am not sure what to make of comparing the structure of a sentence or painting with the structure of conscious states, either in cognitive or neural terms.

Let me also here briefly mention some suggestive empirical evidence that I think supports the HOT theory in general and the WIV in particular. I expand on these themes later in this chapter and in chapter 9.

(1) Gerald Edelman and others have argued that *feedback loops* (or *reentrant pathways* or *back projections*) in the neural circuitry of the brain are essential for conscious awareness (Edelman and Tononi 2000a, 2000b). As Churchland puts it, "The idea is that some neurons carry signals from more peripheral to more central regions, such as from V1 to V2, while others convey more highly processed signals in the reverse direction. . . . It is a general rule of cortical organization that forward-projecting neurons are matched by an equal or greater number of back-projecting neurons" (2002, 148–149). The brain structures involved in loops seem to resemble the structure of at least some form of HOT theory; namely, lower-order and higher-order states are combining to produce conscious states. More specifically, such evidence seems to support the WIV because of the intimate and essential relationship between the "higher" and "lower" areas of the brain involved. There is mutual interaction between the relevant neuronal levels. Edelman and Tononi, for example, emphasize the global nature of conscious states, and it is reasonable to interpret this as the view that conscious states comprise both the higher- and lower-order states. As they describe it, what they call the "dynamic core" is generally "spatially distributed and thus cannot be localized to a single place in the brain" (Edelman and Tononi 2000a, 146). It seems to me that their description fits more naturally with the WIV. It is also difficult to understand the notion of feedback loops in the painting and sentence analogies.

(2) The importance of higher-order *concepts* in conscious experiences is also readily apparent. Part of the reason why Edelman and others believe that back projections play a prominent role in consciousness is that "perception *always* involves classification; conscious seeing is *seeing as*" (Churchland 2002, 149). This is a key aspect of any HOT theory, as we will see in chapter 6 in my defense of the Conceptualism Thesis.

Returning to Levine's objection, then, it is at best misleading to treat the lower- and higher-level parts of a conscious state as potentially bifurcated. Thus while the standard HOT theory appears to have a serious problem with respect to possible misrepresentation, the WIV is better able to handle this objection. Levine's options three and four thus seem particularly open to a defender of the WIV.

4.2.3 More on Targetless HOTs

Let us further examine cases where the HOT has no target at all. Rosenthal frequently refers to confabulation and dental fear as examples of targetless or "hallucinatory" HOTs. Confabulation typically involves (falsely) thinking that one is in an intentional state or, better, making erroneous claims

with regard to the causes of one's intentional states (Nisbett and Wilson 1977). Dental fear occurs when a dental patient seems to experience pain even when nerve damage or local anesthetic makes it impossible for such a pain to occur. Perhaps the patient's fear has been mistaken for pain, but it may also be that the patient has a HOT about being in pain when in fact no pain is present.

What I find most puzzling in this discussion is the implication that we are talking about possible misrepresentation *within* first-order conscious states. Rather, it seems to me that these cases involve fallible *introspection*, and thus misrepresentation at this level is not a problem at all. Both the WIV and HOT theory can and should acknowledge that one might be mistaken when one introspects. When one flounders around for an explanation in the case of confabulation, one seems to be rationalizing about one's own behavior or mental states. This is presumably what occurs during some instances of introspection and results in one (falsely) believing that one has a particular mental state. Confabulation involves a process whereby one is searching, as it were, for an explanation of one's behavior. But since no plausible introspective explanation arises, one tends to make one up, that is, to literally create or cause one instead. Indeed, Rosenthal sometimes refers to "confabulatory introspective awareness" (2005, 125). In short, then, the appearance/reality distinction still applies to introspection, but not within a complex conscious state. I elaborate on this theme in section 4.5.

Much the same applies to the dental patient. Intense and fearful introspection can cause the patient to confuse fear with a pain or represent being in pain when there is no pain. However, it seems to me that another explanation is more plausible. Owing to the fear and expectations of the dental patient, this case is better explained via what Hill (1991) calls "activation." As we saw in the previous chapter, introspection can actually involve the creation of a lower-order conscious state. It might just be that a genuine pain is created "top-down," so to speak, and is thus felt by the patient. I can surely, via introspection, cause myself to have a desire for lasagna if I think about it for a minute or so. In any case, we can happily acknowledge that a *conscious* HOT (= introspection) can either have no target (and thus be fallible) or create a target state (and thus really result in a conscious state). But in neither case does this threaten the WIV. Importantly, however, the main misrepresentation objection raised earlier to Rosenthal's view does not apply in these cases. If we are now referring to fallible *conscious* HOTs (or introspection), then it makes perfect sense that subjects would still subjectively experience those states in an indistinguishable way.

Finally, it is worth briefly acknowledging other related and familiar examples found in the psychological literature on metacognition (Koriat 2007). I have in mind "feeling of knowing" judgments (FOK) and "tip of the tongue" phenomena (TOT). Rosenthal replies that TOT and FOK are examples of being conscious "of knowing something without being conscious of the thing we know" (2000b, 204). I can be conscious *of* the state only *as* a state that carries some information, such as Mark Twain's real name, without being conscious of the information itself. We need not be aware of every aspect of our conscious states. I largely agree with Rosenthal here. Once again, however, TOT and FOK are best understood as introspective states, since one is typically consciously and actively searching for the knowledge in question. Moreover, these cases have more to do with attempts to recall knowledge through memory, as opposed to the typical assertoric HOT.[6]

In closing, then, the charge that HOT theory cannot handle cases of misrepresentation is a serious one. However, authors such as Levine and Neander do not properly explore the options and resources available to the HOT theorist, and they hastily conclude that all versions of HOT theory must choose between arbitrary or untenable alternatives. But the WIV can resolve this problem even if we concede that standard HOT theory cannot. I return to the topic of misrepresentation later in this chapter, as well as in chapter 6, where I further refine my response to the misrepresentation problem.

4.3 The Problem of the Rock

Here is another pair of serious and related objections: HOT theory has been or could be attacked from two apparently opposite directions. On the one hand, we have what Stubenberg (1998) has called "the problem of the rock," which, if successful, would show that HOT theory proves too much. On the other hand, it might also be alleged that HOT theory does not or cannot address the so-called hard problem of phenomenal consciousness. If so, then the HOT theory would prove too little. We might say, then, that HOT theory is arguably between a rock and a hard place. In this and the next section, I critically examine these objections and defend HOT theory (and especially the WIV) against them. In doing so, I will show that it proves neither too little nor too much but is just right. I also show that these two objections are really just two sides of the same coin, and that HOT theory is immune to David Chalmers's (1995, 1996) criticisms of other attempted reductionist accounts of consciousness.

Following Stubenberg (1998), I will call the following classic objection from Alvin Goldman to all higher-order (HO) theories of consciousness "the problem of the rock":

> The idea here is puzzling. How could possession of a meta-state confer subjectivity or feeling on a lower-order state that did not otherwise possess it? Why would being an intentional object or referent of a meta-state confer consciousness on a first-order state? A rock does not become conscious when someone has a belief about it. Why should a first-order psychological state become conscious simply by having a belief about it? (Goldman 1993, 366)

Clearly this objection could be devastating. If any HO theory succeeds, then it would be guilty of proving too much. That is, it would make too many things conscious and would thereby be reduced to absurdity. Thus a HO theorist must block the generalization to rocks and the like without sacrificing an informative analysis of consciousness. Two preliminary points may first be made on behalf of HO theories: (1) Every HO theorist acknowledges that the meta-state does not *always* confer consciousness on its target state. Thus, for example, the HOT must arise in a suitably unmediated way for it to confer consciousness. Recall that a HOT must not arise via inference from observing one's own behavior. (2) The HO theorist might also object to Goldman's use of the term "belief" instead of "thought." For example, one might argue that beliefs are best understood as dispositional states and so are not up to the task of conferring consciousness on anything, including both rocks and mental states. Of course, it is not always easy to distinguish between an *occurrent* belief and a thought,[7] so perhaps the problem of the rock can simply be recast in terms of occurrent belief.

Nonetheless Goldman's analogy is still suspect in at least two ways: (a) As Lycan observes, the difference between rocks and psychological states is simply that psychological states "are themselves mental. . . . It seems psychological states are called 'conscious' states when we are [aware] of them, but nonpsychological things are not" (Lycan 1996, 24). Thus, in Goldman's analogy, we do not first have a mental state that then becomes the object of a meta-state. Instead there is a rock, clearly a "nonpsychological thing," which becomes the object of a mental state (cf. Byrne 1997, 110–111). (b) Goldman does not distinguish between conscious and unconscious meta-states in the way that HOT theorists do. Recall that HOT theory says that what makes a first-order world-directed mental state conscious is the presence of an *unconscious* HOT directed at it, but when the HOT is itself conscious, we then have an *introspective* state. So is the belief about the rock conscious or unconscious? It is difficult to make sense of Goldman's analogy in this context. On the other hand, perhaps Goldman and others can

simply insist that the problem remains either way: neither conscious nor unconscious thoughts about rocks make them conscious.

Thus I still think that we need to take this objection seriously. I am not satisfied with Rosenthal's answer to this problem. Recall the distinction between transitive and intransitive consciousness. We saw in chapter 1 that transitive consciousness is the "conscious of" sense that is typically attributed to subjects, whereas intransitive consciousness is the "is conscious" sense that is often also attributed to individual mental states. Recall also that "a property is intrinsic if something's having it does not consist, even in part, in that thing's bearing some relation to something else" (Rosenthal 1997, 736). He then replies to the problem of the rock as follows:

> Being transitively conscious of a mental state does in a sense make it intransitively conscious. But that is not because being conscious of a mental state causes that state to have the property of being intransitively conscious; rather, it is because a mental state's being intransitively conscious simply consists in one's being transitively conscious of it. The mistake here is to suppose that a state's being intransitively conscious is an intrinsic property of that state. If it were, then being intransitively conscious could not consist in one's being transitively conscious of being in that state unless being thus conscious induced a change in that state's intrinsic properties. This objection is at bottom just a disguised version of the doctrine that being intransitively conscious is an intrinsic property. (738–739)

It is, to be sure, important not to construe Rosenthal's HOT theory as holding that being "conscious of" a mental state *causes* that state to have the property of being intransitively conscious. For Rosenthal, a mental state's being intransitively conscious simply *consists in* one's being transitively conscious of it. The former erroneous causal reading of Rosenthal's HOT theory may indeed contribute to the intuitive force of the rock problem.

Nonetheless it is first not clear to me exactly why Rosenthal thinks that the objection is "just a disguised version of the doctrine that being intransitively conscious is an intrinsic property" of mental states. I can easily imagine someone sympathetic to his view that HOTs are extrinsic to the states rendered conscious also being troubled by this problem. Indeed, Goldman's initial formulation of the problem seems quite explicitly to separate the meta-state from the lower-order state. But, once again, Goldman ultimately wants to know exactly *how* or *why* such meta-states can "confer" consciousness on their objects. As Byrne (1997, 110) notes, at worst the belief "that intransitive consciousness is intrinsic is a plausible *consequence* of the objection, not the basis of it."

Second, Rosenthal accuses the proponents of this objection of believing, like Descartes, in the intrinsic nature of consciousness. But, as we have

already seen, he also mistakenly seems to think it follows from this that consciousness would then be essential to all mental states and thus unanalyzable (or "simple").[8] Rosenthal is so concerned to avoid the notion that consciousness is an intrinsic property of any mental state that he overlooks the possibility that conscious mental states might consist of *both* the lower-order state and the HOT. But if Rosenthal adopts this view, then he has two problems: (a) the HOT, or "conscious making property," would be intrinsic to the conscious state, which is clearly at odds with Rosenthal's own considered view; and (b) Rosenthal would then not be able simply to dismiss proponents of the problem of the rock as those who just mistakenly treat consciousness as an intrinsic property of intransitively conscious mental states.

In any case, I think that Stubenberg is right when he says that what motivates the proponent of the problem of the rock is the worry that the relation *being the target of a higher-order representation* is the wrong relation. The worry is fueled by one's inability to comprehend how entering into this relation is supposed to promote an unconscious state to consciousness. Those who raise this objection are not just "begging the question against Rosenthal by simply restating their intrinsicalist creed" (Stubenberg 1998, 195). However, unlike Stubenberg, I do not think that the problem of the rock should cause one to give up on some version of a HO theory of consciousness.

So where do we go from here? It is first necessary to return to Lycan's response to the problem. Although he did not go far enough, Lycan does take the first crucial step. We must first and foremost distinguish rocks and other nonpsychological things from the psychological states that HO theories are attempting to explain. HO theories must maintain that there is not only something special about the meta-state (as we will see in the next section) but also something special about the *object* of the meta-state, both of which, when combined in certain ways, result in a conscious mental state. The HO theorist must initially boldly answer the problem of the rock in this way to avoid the *reductio* whereby a thought about *any x* will result in *x*'s being conscious. So the HOT theory does not really prove too much in this sense, and various principled restrictions can be placed on the nature of both the lower-order state and the meta-state to produce the mature theory. In this case, a rock is not a mental state, and so having a thought about a rock will not render it conscious. After all, the HOT theory is attempting to explain *what makes a mental state* a conscious mental state. This is not properly recognized by those who put forward the problem of the rock.

Two further moves can also be made. First, recall from chapter 2 that it might be wise to raise the next natural question: what makes a state a *mental* state? As we have seen, there are differing views here. One might, for example, insist that mental states must fill an appropriate causal-functional role in an organism. Alternatively, one might even simply identify mental states with certain neural or biochemical processes in an organism (Crick 1994). Either way, however, it is clear that external objects, such as rocks, cannot meet these criteria. The LO states in question thus have certain special properties that make it the case that they become conscious when targeted by an appropriate HOT. It is also important to note that this response effectively handles other related objections to the HOT theory. Various *internal* states such as cancer (Dretske 1995, 97, 100) and liver states (Block 1995, 280) are also ruled out by these criteria.[9]

Second, in a similar vein, if we return to the idea that the meta-state is an intrinsic part of a complex conscious state, then it is also clear that rocks cannot be rendered conscious by the appropriate HOT or MET. This is because, in such a view, the MET must be more intimately connected with its object, and it is most natural to suppose that the target object must therefore be "in the head." That is, both "parts" of the complex conscious state must clearly be internal to the organism's mind. Van Gulick (2000, 2004), who calls this "the generality problem," makes a similar point when he says that "having a thought . . . about a non-mental item such as the lamp on my desk does not make the lamp conscious . . . because [the lamp] cannot become a constituent of any such global [brain] state" (Van Gulick 2000, 301). Thus it is difficult to compare the inner (mental)/inner (mental) relation as described by HO theories to the inner (mental)/outer (rock) relation described in Goldman's initial objection. Like the WIV, this move provides Van Gulick and any intrinsic HO theorist with an additional counterargument not available to standard HOT theory.[10]

The problem remains, however, that this reply to the problem of the rock still invites a natural response, as we saw earlier. Such replies typically come in the form of "why questions" or demands for further explanation: *why* or *how* can such meta-states confer consciousness on their mental objects? Recall also part of the earlier quote from Stubenberg: "The worry is fueled by one's inability to comprehend *how* entering into this relation is supposed to promote an unconscious state to consciousness" (1998, 195; italics mine). However, it is interesting to note that we now have a somewhat different, though importantly related, objection. Instead of claiming that HO theories prove too much, a question is now raised about how or if they can explain anything at all regarding state consciousness. That is, the

question has shifted from arguing that HO theories (if true) are too strong in the sense of making too many things conscious to objecting that they are much too weak to explain phenomenal consciousness. It is, in some ways, a curious counterreply, given the initial problem of the rock. This shift to a demand for a satisfying explanation will sound familiar to those aware of what David Chalmers (1995) calls the "hard problem" of consciousness. It is now time to turn our attention to that problem before addressing some additional objections to the WIV, but one key point here is that the problem of the rock and the hard problem are really just two sides of the same coin.

4.4 The Hard Problem of Consciousness

4.4.1 The Challenge

David Chalmers presents the hard problem of consciousness as follows:[11]

> The really hard problem is the problem of *experience*. . . . A subjective aspect. . . . It is undeniable that some organisms are subjects of experience. But the question of *how* it is that [organisms] are subjects of experience is perplexing. *Why* is it that when our cognitive systems engage in visual and auditory information-processing, we have a visual or auditory experience[?] . . . *How* can we explain *why* there is something it is like to entertain a mental image, or to experience an emotion? It is widely agreed that experience arises from a physical basis, but we have no good explanation of *why* and *how* it so arises. (1995, 201; italics mine after the first instance)

The similarity between parts of this quote and any reasonable counterreply to the problem of the rock should be obvious. The number of "why" and "how" questions in Chalmers's quote echoes the demand for explanation that we saw at the end of the last section. Let me be clear about the connection between the initial problem of the rock and the hard problem. One way to put the dialectic thus far is as follows. Problem of the rock: "If HO theory is true, why don't some objects of thoughts (such as rocks) become conscious when thoughts are directed at them?" HO theorist: "Well, rocks aren't mental states that are in the head and thus have importantly different properties than rocks, and so on." Counterreply (hard problem): "But exactly how or why does having HOTs directed at mental states make them conscious or explain conscious experience?"

To be sure, a version of this last question can even be found in Goldman's original objection (and is even clearer in the Stubenberg quotation), but this is precisely why the two objections are two sides of the same coin. Both objections rest on a challenge about what is needed for a theory of consciousness to render consciousness intelligible. Nonetheless there is the subtle shift from addressing the nature of the *objects* of thoughts to the

question of how or if invoking HOTs can provide a satisfying explanation of state consciousness. Moreover, there is the move from objecting that the HOT theory, if true, is too strong, at least in the sense that it would make too many things conscious, to the complaint that the HOT theory, if true, does not present a strong enough explanation of phenomenal consciousness.

In his 1995 paper, Chalmers is not mainly concerned with HO theories of consciousness; indeed, HO theories are noticeably absent from his treatment. He is instead concerned to show how various other attempts to explain consciousness have failed or merely address the relatively "easy problems" of consciousness, such as explaining the integration of information by a cognitive system or the ability of a cognitive system to access its own internal states. Thus the bottom-line objection is once again that various theories of consciousness do not explain enough.

We must first acknowledge that any HO theorist should be willing to take up this challenge. We must face the hard problem head on, though we may never fully satisfy those who advance it. Any theory of consciousness should attempt to explain the "what it is like" of conscious experience. This has perhaps not always been the case. Stubenberg (1998), for example, alleges that HO theorists have intentionally avoided this type of problem or considered it to be unimportant. He argues that HO theories do not explain or even address *the nature* of qualitative experience. Stubenberg uses HO theorists' own words against them. For example, Lycan says that "I am not here addressing issues of qualia or phenomenal character. . . . There may be Inner Sense theorists who believe that their views solve problems of qualia; I make no such claim, for I think qualia problems and the nature of conscious awareness are mutually independent" (1996, 15). Stubenberg also cites Rosenthal's view that "the properties of being conscious and having sensory quality are independent of one another, and a satisfactory account of each property requires us to investigate them separately" (1991, 16).

While Lycan and Rosenthal may seem to contradict such remarks elsewhere,[12] I agree with Stubenberg that many HO theorists have been slow to address the hard problem and have often avoided it intentionally.[13] However, I also disagree with him in one important way: as I will argue hereafter, HOT theory and the WIV can explain qualitative experience even if, as we have seen, Rosenthal argues for the independence of consciousness and sensory qualities. For one thing, it surely does not follow that when the sensory qualities *are* conscious, they cannot be explained in terms of the HOT theory.

4.4.2 A Proposed Solution

Before offering an answer to the hard problem, it is crucial to recall from chapter 2 that some theories of consciousness are entirely nonreductionist. However, HOT theory is reductionist in the sense that state consciousness is to be explained in terms of something else, that is, unconscious mental states. It is also important to remember that HOT theory does not attempt to explain consciousness in nonmentalistic or naturalist terms. Of course, HOT theorists, including myself, tend to be materialists in the end but prefer to leave that empirical question for a separate second-step reduction to be filled in later by brain science.

I believe that HOT theory can provide necessary and sufficient conditions for what makes a mental state conscious, but whatever realizes that theory *in our brains* is a separate empirical question. On the other hand, other theories of consciousness are reductionist in the stronger sense that they attempt to explain consciousness *directly* in physical or neurophysiological terms (e.g., Crick and Koch 1990; Crick 1994). Chalmers is well aware of these distinctions, but he argues that all such attempts still fail to address the hard problem. The earlier quotation from Chalmers does show, however, how statements of the hard problem can sometimes gloss over these differences. Nonetheless the key common general question is: how exactly does x (where x is invoked by some alleged theory of consciousness) explain how consciousness arises from the presence of x? Recall that we encountered a somewhat similar question in connection with the problem of the rock.

According to the HOT theory, the x is to be filled in with "the presence of a suitable higher-order thought." But then the hard problem asks: *why* or *how* exactly does the presence of such a HOT result in a conscious mental state? Some of what I have said in Gennaro 1996 was at least an indirect attempt to answer this question. However, at that time, I did not explicitly have Chalmers's hard problem in mind. I want to show now how that basic idea can be brought to bear on this problem.

The solution is that HOTs (or METs) explain how consciousness arises because the *concepts* that figure into the HOTs are presupposed in conscious experience. Let us stick to first-order perceptual states. Very much in a Kantian spirit, the idea is that we first passively receive information via our senses. This occurs in what Kant (1781/1965) calls our "faculty of sensibility," which we might think of as early perceptual processing. Some of this information will then rise to the level of unconscious mental states, which can also cause our behavior in various ways. But such mental states do not become conscious until the faculty of understanding operates on them via

the application of concepts. I contend that we should understand such concept application in terms of HOTs (or METs) directed at the incoming information. Thus I consciously experience the brown tree *as a brown tree* partly because I apply the concepts "brown" and "tree" (in my HOTs) to the incoming information via my visual perceptual apparatus. More specifically, I have a HOT such as "I am seeing a brown tree now." Kant urges that it takes the cooperation of both the sensibility and understanding to produce conscious experience. Regarding the sensibility and understanding: "Objects are *given* to us by means of sensibility. . . . They are *thought* through the understanding, and from the understanding arise concepts" (A19/B33).[14]

It is crucial to remember that these HOTs (or METs) are not themselves conscious, and so HOTs and their concepts are "presupposed" in conscious experience. We might say that the understanding unconsciously "synthesizes" the raw data of experience.[15] Recall that unconscious mental states also involve some form of conceptualization or categorization insofar as they have intentional content. However, part of the motivation for HOT theory is to explain when and how an unconscious state becomes conscious, and the answer is that the subject becomes aware of the state; that is, a HOT or MET is directed at it (recall the Transitivity Principle). Thus the concepts in question must also be in the HOTs, and they are primarily responsible for the "what it is like" nature of qualitative experience.

Rosenthal has also argued in a somewhat similar Kantian fashion. Recall his well-known example of wine tasting. He first says, "Learning new concepts for our experiences of the gustatory and olfactory properties of wines typically leads to our being conscious of more fine-grained differences among the qualities of our sensory states. Similarly with other sensory modalities . . . new concepts appear to generate new conscious sensory qualities" (2002a, 413). Rosenthal then uses the example of hearing the sound of an oboe. He argues that if we systematically remove *all* the relevant concepts (e.g., "sound of a woodwind") involved in having that conscious experience, then there would no longer be the conscious experience. This seems true for any conscious experience. As Rosenthal puts it, it is "plausible that peeling away that weakest HOT would result, finally, in its no longer being like anything at all to have that sensation" (2002a, 414). The concepts that we have clearly color the very experiences we have, and removing all of them would eliminate the experience itself. Indeed, according to the HOT theory, having such concepts is both necessary and sufficient for having subjective conscious experience.

Now I submit that this is an adequate answer, or at least an initial answer, to the challenge at hand.[16] Of course, some will no doubt always remain dissatisfied and will want to ask a further question: *why* does the higher-order application of concepts give rise to conscious experience? But this, I suggest, is not a legitimate question. We have already reached the rock-bottom brute fact about the way that conscious minds work, and the chain of explanation has already come to an end. The Kantian idea that concepts make our experience of the world possible is a widely held and plausible view about the nature of conscious experience. I do not think that it makes sense to ask why this is so. As Strawson puts it: "[If] any item is even to enter our conscious experience we must be able to *classify* it in some way, to *recognize* it as possessing some general characteristics" (1966, 20; italics mine).[17] Notice that this solution is unlike reductionist accounts in *nonmentalistic* terms and so is immune to Chalmers's criticism about the plausibility of those theories. For example, there is no problem about how a specific *brain activity* produces conscious experience. Chalmers's criticism that *functional* explanations are inadequate because one can always ask, "*Why is the performance of these functions accompanied by experience?*" (1995, 203), is equally beside the point. HOT theory is not a functional explanation that merely addresses the easy problems of consciousness. In any case, HOT theory contends that a reductionist theory of consciousness can be provided in mentalistic terms in a way that can solve the hard problem. This is also, then, the answer to those who demand a further explanation as a counterreply to the problem of the rock.

There are, no doubt, other questions regarding concepts that do make sense and need to be answered: What are concepts? What is it to possess a concept? How do we apply concepts? How do we acquire concepts? Which, if any, concepts are innate? These are notoriously difficult questions to answer, each with a long history of failed or questionable attempts. For example, the search for necessary and sufficient conditions of *application* for any given concept often seems doomed to failure. Analyzing concept possession in terms of, say, mental images or mere behavioral discrimination seems insufficient for various well-known reasons. This chapter is not the place to put forth a complete theory of concepts.[18] My point here concerning HOT theory is simply that these legitimate questions are far different in structure from the earlier illegitimate question raised in relation to the hard problem, that is, *why* does the higher-order application of concepts give rise to conscious experience? In other words, it may be extremely difficult to explain how we acquire or apply concepts, but that differs from demanding a further explanation for the fact that, *given the presence of concepts (in*

the HOTs), conscious experience results. If there is a real hard problem of consciousness, I suggest that it has more to do with concept acquisition and application.

As I mentioned earlier, HOT theory is notoriously absent from Chalmers's 1995 discussion, and it is only briefly mentioned in his 1996 book. Perhaps he would place the HOT theory in the same category as others who approach the problem of consciousness at the level of cognitive psychology. For example, Baars's (1988) global workspace theory of consciousness and Dennett's (1991) multiple-drafts model could be viewed as similar to the HOT theory in some respects. But Chalmers urges us to agree that neither of these theories really addresses the hard problem. Baars leaves unanswered the question: why is the information in the global workspace experienced? And Dennett's theory "is largely directed at explaining the reportability of certain mental contents" (Chalmers 1995, 205). Thus such theories never really do address the hard problem. However, even if Chalmers is right in his criticism of Baars and Dennett, I have tried to show how HOT theory is in a better position to address the hard problem. HOT theory is not merely dealing with an easy problem (e.g., in terms of a purely functional relation); nor is it ignoring the hard problem altogether. It is attempting to explain the very *structure* of conscious mental states, as well as *how* such states come to have their qualitative feel.

Chalmers makes several remarks akin to HOT theory in presenting what he calls "naturalistic dualism," which is the view that experience is a basic or fundamental feature of the universe over and above the properties invoked by physics. For example, when presenting his "principle of structural coherence," Chalmers (1995) describes a close connection between "consciousness" and "awareness" (cf. Chalmers 1996, 218–29). In his response to various authors (in Shear 1997), Chalmers also says that "I find it plausible that there is an intimate relationship between consciousness and thought" (396). It is clear that this is a view shared by HOT theorists. A higher-order thought (or awareness) of a lower-order state renders that state conscious. I am not suggesting that Chalmers is a closet HOT theorist, but such quotes reveal an important agreement between him and HOT theory. Instead Chalmers ultimately uses the notion of "information" in defending a "double-aspect" theory whereby information "has two basic aspects, a physical aspect and a phenomenal aspect" (1995, 216).

4.4.3 Conceivability, Necessity, and the HOT Theory

I do not want to give the impression that I agree with well-known "conceivability" and "zombie" arguments against physicalism. The idea again

is basically that if it is conceptually possible for there to be a physically identical creature to me that lacks conscious experience, then ultimately materialism is false (recall the discussion from section 2.2.2). My main point here is that HOT theory and the WIV are in better shape to deal with such arguments and can even avoid them entirely. As a matter of fact, I am quite sympathetic to those who have responded in a variety of ways to Chalmers's use of such arguments.[19] In particular, I do not understand how we can draw any ontological conclusion about the actual world from such arguments. I have little to add to the many standard criticisms of the conceivability argument, and in any case, the topic goes beyond the scope of this work.

Still, Chalmers might object to my solution by asking, "How can the HOT theory express a necessary truth, especially since its denial does not seem to involve a contradiction?" "Even if consciousness necessarily presupposes concepts, why must those concepts be constituents of *higher-order* thoughts?"[20] Surely, he might say, it is *possible* for the HOT theory to be false, and if so, then the rest of his argument can proceed from there. This line of thought is, once again, clear from the earlier passage: "The conceivability of zombies seems . . . obvious to me. . . . While this possibility is probably empirically impossible, it certainly seems that a coherent situation is described; I can discern *no contradiction* in the description" (1996, 96; italics mine).

It seems to me that at least the following three responses are open to the HOT theorist. First, to the extent that Chalmers is willing to allow "brute intuition" to guide possibility, then I must admit that I do find it extremely difficult to conceive of a creature with conscious experience but without HOTs, or vice versa. This is based on my critical and substantial reflection on the nature of conscious mental states, which should count for something. More importantly, however, even if I do admit the possibility of consciousness without HOTs (or vice versa), it would seem that something similar is also true of Chalmers's own view. To the extent that Chalmers is merely engaged in conceptual analysis, surely it is not *contradictory* to suppose that, for example, consciousness is not a fundamental ontological feature of the world. Surely it is also not *contradictory* to suppose that consciousness is not tied to "information" in the way that Chalmers speculates. Surely, then, there are possible worlds where Chalmers's view is false. Perhaps he would deny such a possibility, but then I think I can do the same with respect to the HOT theory. On the other hand, if Chalmers accepts this possibility, then presumably he would not take it as proving that his ontological view about the actual world is false. And so I should again be

permitted to say the same about the HOT theory. If that is good enough for Chalmers, then it is good enough for me. It may be that Chalmers is more concerned with the *ontology* of consciousness in this context instead of mere conceptual analysis, and so he might object to treating his view as mere conceptual analysis. But then again, the same could go for proponents of the HOT theory, as well as for various naturalist theories. In all these views, including Chalmers's own, we should then wonder how conceiving of their falsity can force us to conclude that they are false as an ontological view about the actual world.

A second, more provocative reply is much more involved. It requires us first to define an often used, but sometimes conflated, threefold distinction. There are analytic and synthetic *statements* or *sentences*; the former we can define as a sentence whose denial is contradictory, whereas the latter are simply sentences that are not analytic. These terms are usually understood as making a claim about the *meaning* of various concepts involved in sentences, as in the classic analytic example "all bachelors are unmarried." There is also a priori and a posteriori (or empirical) *knowledge*: the former can be known independently of sense experience whereas the latter is acquired via the senses. Finally, there are contingent and necessary *truths*; the former are true in this and (perhaps) some other possible worlds, and the latter are true in all possible worlds.

The problem with the objection, then, is that it assumes that only analytic statements can express necessary truths. Now, first of all, the familiar "water is H_2O" example shows that this assumption is false (Kripke 1972). Of course, Chalmers and others know this very well, but they then typically argue at great length about the alleged differences between the water–H_2O case and consciousness. Indeed, this is the main purpose behind Chalmers's two-dimensional semantics where he distinguishes between the primary and secondary intensions (meanings) of a concept. However, what really causes trouble for his opponents here is that "water is H_2O" is known a posteriori.[21] After all, Chalmers makes it clear that his main targets are the proposed solutions to the hard problem that mention neurophysiological properties. This is what Botterell (2001, 22–23) aptly calls "*a posteriori* physicalism" and then quotes Chalmers as having in mind the view that psychological nature is "not necessitated a priori by physical [nature], but . . . [is] necessitated a posteriori by physical [nature]" (Chalmers 1999, 474). Botterell rightly characterizes this view as follows: "In short, a posteriori physicalism maintains that although there is a *necessary* entailment of psychological nature by physical nature, there is no *a priori* or *conceptual* connection between the two" (2001, 23).

My main point here, however, is that whatever force Chalmers's views have in relation to a posteriori physicalism, they do not apply to the HOT theory. Unlike a posteriori physicalism, the HOT theory does hold that the relationship between the explanandum and explanans is a priori and, in this sense, a conceptual one. A *logical entailment* holds between the content of the HOT theory and statements about the nature of conscious experience, and Chalmers agrees that "if B-properties are logically supervenient on A-properties according to primary intensions, then the implication from A-facts to B-facts will be a priori" (1996, 70). This interpretation of HOT theory is also supported by Van Gulick when he says that it proposes to offer "logically sufficient condition[s] from which one could deduce the existence and nature of the relevant feature of consciousness" (1995c, 70). Much of my discussion of the Transitivity Principle (in sec. 2.4.1) can be understood as an example of the a priori reasoning at the heart of HOT theory.

Thus even if I am willing to concede that *the statement* of the HOT theory is synthetic simply because its denial does not seem to result in an explicit contradiction, we can still reply as follows: if there has ever been any dispute about whether or not synthetic a posteriori truths can be necessary, it is normally because they are known *a posteriori*. But if the fundamental thesis of the HOT theory (i.e., the Transitivity Principle) is known a priori, then, even according to Chalmers, the idea of an a priori necessary truth would be far less controversial. HOT theory can therefore embody a synthetic, a priori, and necessary truth. If one agrees with Kant (as I do) that there are synthetic a priori truths (such as mathematical truths and "every event has a cause"), then the HOT theory can again avoid Chalmers's arguments against a posteriori physicalism. Logical entailments can be found in the HOT theory in a way that is absent in purely physicalist theories of consciousness. Thus even if the *statement* of the HOT theory is synthetic, we can still maintain that it is known a priori and *expresses* a necessary truth. At the least, then, Chalmers is unjustly ruling out synthetic a priori truths, though he rightly recognizes the close link between truths known a priori and their metaphysical necessity.

It may seem odd to talk about the HOT theory as a synthetic statement together with the notion that there are logical entailments between it and statements about conscious experience. But there is really no problem here. Suppose Kant was right about the synthetic nature of statements in geometry. It would still make perfect sense to talk about logical entailments *between* such statements and other statements in geometry or even in other disciplines. Analogously, even if the main claim of the HOT theory is synthetic, we can still make sense of logical entailments between a particular

example of the HOT theory at work and the corresponding statement about the nature of that particular conscious experience.

Third, a HOT theorist might even agree with Quine (1951) and question the very distinction between analytic and synthetic statements. At the least, it could even be argued that some apparently synthetic statements are "really" analytic in the sense that, to use Kant's phrase, they contain predicates that are *implicitly* contained in the subject. That is, perhaps one can admit that the denial of a sentence does not lead to an *explicit* or *transparent* contradiction, but it nonetheless does eventually lead to an *implicit* one. This line is not always so easy to draw. After all, even clear classic cases such as "all bachelors are unmarried" involve some unpacking of the notion of "bachelor." Thus a HOT theorist may even wish to maintain that, say, "being accompanied by a HOT" is implicitly contained in the *concept* "conscious mental state." As Chalmers himself rightly points out: "In certain cases, the decision about what a concept refers to in the actual world involves a large amount of reflection about what is the most reasonable thing to say" (1996, 58). Chalmers then goes on to admit that such a process is still best understood as engaging in a priori reasoning (58–59). In the end, however, what is most important to remember is that we are identifying the two *properties* in question. Analyzing the *concepts* is one thing, but it is the ontological identity between *properties* that is most central.

4.4.4 First Application: Van Gulick's Approach

It will be helpful to place this discussion in the context of a paper by Robert Van Gulick (1995c). He asks the metaphilosophical question "What would count as explaining consciousness?" and his purpose is to "untangle and clarify the various distinct issues which sometimes get run together" (61). He accomplishes this by distinguishing

(A) various *explananda* of consciousness (i.e., what needs to be explained);

(B) the domain of *explanans* (i.e., in what terms we might construct an explanation); and

(C) a *linking relation* that must hold between the explanandum and the explanans to provide a satisfactory explanation.

Under (A) Van Gulick lists five features: (A1) the un/conscious state distinction, (A2) non/conscious creature distinction, (A3) qualitative aspect, (A4) phenomenal aspect, and (A5) subjectivity.

Under (B) he lists the following: (B1) physical, (B2) functional, (B3) naturalistic, (B4) nonconscious mental states.

And under (C) he mentions (C1) logical sufficiency, (C2) nomic sufficiency, (C3) intuitive sufficiency, and (C4) predictive and pragmatically useful modeling.

Van Gulick points out how this menu generates ninety-six possible interpretations of his original ambiguous question. Furthermore, he argues that some theories are clearly only concerned with one or two features from each category. For example, a hard-core reductionist approach would presumably only use B1 or B3 and probably C2 (or perhaps C4) to explain one or more items under (A). Although Van Gulick does not demand that a theory of consciousness explain all the explananda he lists under (A), it seems reasonable to hold that any good theory of consciousness should be able to do so. Presumably, this is precisely the force behind the hard problem. In particular, a theory of consciousness should be able to explain A3 through A5 (qualitative, phenomenal, subjectivity). Although Van Gulick makes several subtle and somewhat terminological distinctions between those three explananda, he acknowledges that they could all be grouped under the general heading of "first-person qualitative experience."

Against this backdrop, I wish to make several points regarding the HOT theory and the hard problem. First, it is interesting to note that (A) differs from (B) and (C) in one important way; namely, all of A's features should be explained by any good theory of consciousness, whereas we are not required to pick more than one item from either of the other two categories. Indeed, for example, using B1 as the explanans (physical) would even rule out using B4 (nonconscious mental states).

Second, Van Gulick rightly notes that HOT theory is probably most often understood as explaining A1, that is, the un/conscious state distinction. HOT theory can also presumably handle A2 (non/conscious creature distinction) fairly well. However, as I have argued, we can and should also explicitly take up the challenge of explaining A3 through A5, and after all, part of the goal of this entire book is to show how a version of HOT theory can explain qualitative consciousness.

Third, in the domain of explanans, HOT theory clearly relies on a B4 approach; that is, explaining consciousness in terms of unconscious mental states. Again, there is no requirement for a theory of consciousness to use more than one explanans from the B list to explain consciousness. Indeed, it is crucial to note that Van Gulick separates B4 from both functional and physical explanations (though they could all be used together). As I argued

earlier, then, HOT theory is immune to the criticisms that Chalmers levels against these two approaches.

Fourth, Van Gulick associates HOT theory with C1; that is, the linking relation of logical sufficiency. I agree with this, as should be clear from my earlier remarks. Once again, recall that Van Gulick rightly states that HOT theory does propose to offer "logically sufficient condition[s] from which one could deduce the existence and nature of the relevant feature of consciousness" (1995c, 70). Indeed, I have argued that HOT theory provides necessary and sufficient conditions for a mental state to be conscious. HOT theory expresses a necessary truth about what makes a mental state conscious (which we can come to know a priori). There *could not* be a conscious mental state that is presented to its owner devoid of *all* conceptualization. A conscious state *must* be presented to its owner in some way or other; that is, it must be thought of under some mode of presentation. So again, then, if HOT theory is defensible, it is immune to Chalmers's arguments against physicalist approaches, particularly when he argues, via zombie arguments, that such a physicalist explanans does not *logically entail* the presence of conscious experience.

Fifth, a HOT theorist can also appeal to the linking relation that Van Gulick calls "intuitive sufficiency" (C3). There is, I believe, something uniquely simple and intuitive about HOT theory's ability to explain consciousness. Of course, as Van Gulick recognizes, "the problem [here] . . . is what strikes us as intuitive is highly context sensitive and relative" (1995c, 71). I am normally not one to rely too heavily on philosophical intuition. Nonetheless I do think that HOT theory has an intuitive appeal over other accounts for the very reasons we saw in chapter 2 in connection with the Transitivity Principle. Moreover, HOT theory preserves these intuitions while avoiding the problems with purely physicalist accounts, which, Chalmers argues, cannot properly respond to the intuition that for any brain process "it is conceptually coherent that it could be instantiated in the absence of experience" (1995, 208).

4.4.5 Second Application: Schröder's Argument

Jürgen Schröder (2001) argues that reductionist accounts of consciousness in physical or "naturalist" terms have greater explanatory power than HOT theories. However, if my argument thus far has been successful, there is one key area that has been ignored by Schröder, namely, that HO theories can provide a better explanation of the logically sufficient conditions of consciousness. That is, unlike naturalist accounts, HOT theory can avoid the concern, raised by Chalmers and others, about the logical relationship

between explanandum and explanans. I suggest that this is a key advantage over naturalist accounts.

Schröder does address one related issue. He responds to the potential criticism that HO theories can provide *necessary* conditions, whereas naturalist accounts cannot, because HO theories allow for the multiple realizability of conscious states in a way that naturalist accounts cannot. For example, if a naturalist theory is couched in neurophysiological terms, then it apparently cannot offer a necessary condition for consciousness because it seems possible for other organisms to have conscious states and yet not have the corresponding neurophysiology. Schröder responds, in part, by suggesting that, say, if a non-HO theory "identifies consciousness with the stability of activation vectors, this property—stability of activation—can be regarded as a relatively high-level property because nothing is determined about the kinds of units that are activated or about the specific form of energy that a state of activation is based on" (2001, 29). Although I doubt that his reply is fully adequate in the end, I do not wish to debate Schröder on this point here. I raise it only to point out an important contrast.

Let us even grant that Schröder can answer that criticism. Unfortunately, he would still not be addressing the key advantage of HO theories raised in the previous subsection, namely, *logical sufficiency from* explanans *to* explanandum. Schröder's response to the foregoing criticism about multiple realizability deals with a *necessary* condition of the explanandum. The direction of explanation here, however, is quite different. Schröder is responding to a concern about the relation *from explanandum to explanans*, whereas in this section we have been more concerned with the (logically sufficient) relation *from explanans to explanandum*. I do not mean to suggest that Schröder intended to address this issue and has failed to do so successfully. He simply never raises it at all. My point, then, is that if I am right, he has altogether ignored one important further advantage of HO theories over naturalist accounts. The HOT theory, especially when viewed as a *philosophical* theory, is preferable to naturalist accounts of consciousness.

I conclude here that we can successfully defend HOT theory, and especially the WIV, against both the problem of the rock and the charge that it does not address the hard problem of consciousness. The HOT theory is therefore not really between a rock and a hard place. The HOT theory does not claim that a metastate about *anything* will render it conscious: the target of the HOT must be a suitable mental state. Furthermore, analyzing the HOT theory in terms of concept application reveals that it can address the hard problem. One advantage of this approach is that HOT theory is immune to the criticisms that Chalmers levels against a posteriori

physicalism. It turns out, though, that these two problems are really two sides of the same coin. The charge of not answering the hard problem is really just a typical counterreply to the problem of the rock.

4.5 Objections and Replies

Now that we have a more fully developed theory and have addressed the hard problem, let us return to the WIV in particular. For the sake of further defending and clarifying the WIV, I will consider a number of objections. Important details emerge throughout this section.

4.5.1 The Unconscious Parts Objection

One might first object that the notion of an unconscious part of a complex conscious state is difficult to understand. What does it even mean to say that state-parts can be conscious or unconscious? The WIV is implausible because the notion of an unconscious part of a conscious state sounds contradictory.[22]

There are at least two replies. First, the objection could be taken as committing the well-known "fallacy of division," namely, that what is true of the whole must be true of each part of the whole. Water extinguishes fire, but oxygen does not. If this is the motivation behind the objection, then it clearly fails. Indeed, if we think of it in a physicalistic way, then such logic would seem to lead automatically to panpsychism. Why should we suppose that *each* part of a complex conscious state is itself conscious? I see no reason to assume that one need be consciously aware of each intrinsic part of a conscious state. Are we consciously aware of everything involved in a conscious state? I suggest not, and this is particularly clear if we think of conscious states as globally represented brain states. All kinds of unconscious mental activity are involved in a conscious state. Although I disagree with Colin McGinn's "mysterian" views on consciousness, he is right when he speaks of there being a "hidden structure of consciousness" such that there are "surface properties, which are accessible to the subject introspectively; and deep properties, which are not so accessible" (McGinn 1991, 111). Indeed, in the WIV, the MET is precisely such a deep property but can surely still be part of (or intrinsic to) the entire complex conscious state. As McGinn puts it: "The subject is not conscious *of* the deeper layer . . . but it does not follow that this layer does not belong intrinsically to the conscious state itself. Just as F can be an intrinsic property of a perceptible object x without being a perceptible property of x, so conscious states can have intrinsic properties that they do not have consciously" (1991, 98).

Second, we must again distinguish between the first-person and third-person perspectives on conscious states. From the first-person point of view, we cannot expect to be consciously aware of all that is "presupposed," to use a Kantian term, in a conscious state. As I have explained previously, we receive information via our senses in our faculty of sensibility, some of which rises to the level of unconscious mental states. But such mental states do not become conscious until the more cognitive "faculty of understanding" operates on them via the unconscious application of higher-order concepts in the METs. Thus I consciously experience the blue wall *as a blue wall* partly because I have the MET "I am seeing a blue wall now." However, it is crucial to note that neither the MET nor the concept application is itself conscious. The understanding unconsciously "synthesizes" the raw data of experience to produce the resulting conscious state. As Kant understood well, there must be significant unconscious (synthesizing) activity implicit in each conscious state. We are not conscious of that activity itself although it is intrinsic and essential to the resulting conscious state. Indeed, it is the MET that makes the state conscious because of the conceptual activity directed at the lower-order mental state.

There is nothing implausible about the existence of unconscious parts of conscious states. On the contrary, such a view is crucial to appreciating the subtlety of the WIV.[23]

4.5.2 Rosenthal's Objection

David Rosenthal (1993b) objects that HOTs (or METs) cannot be intrinsic to conscious states because it would be contradictory or incoherent for a single state to be, for example, a conscious doubt that it is raining and an affirmative (i.e., assertoric) thought that I am in that state. That is, we normally *individuate* states in terms of mental attitude, and we never talk about one state with two attitudes, especially when there are somewhat opposing attitudes desiring and believing that p or doubting and thinking that p.

To reply, let us make an important threefold distinction. First, we have the conscious *state*, that is, the *vehicle* that is identical with a mental representation and is presumably a brain state of some kind. Second, we have the representational *content* of the state in question; that is, what the state is about or directed at. Third, we have the mental *attitude* (or "mode") of the state; that is, what type of mental state it is, for example, a doubt, a thought, a perception, and so on. This threefold distinction is particularly crucial when teasing apart the sometimes subtle differences between the views under consideration.[24]

Unlike standard HOT theory, virtually any form of intrinsic theory will hold that a single vehicle or state can have dual representational *content*. More specifically, the WIV says that the higher-order content is represented in the same state as the first-order content. More to the point raised by Rosenthal, it also seems perfectly possible to have a single conscious state involving two *attitudes*, because one will be directed at its first-order content, and the other will be directed toward its higher-order content. It is not as if a subject is thinking p and not-p or perceiving O but thinking that one does not perceive O. Thus, in the WIV, a complex conscious state, CMS, can have one attitude (a doubt) directed at the weather and another attitude (an assertoric thought) directed at the doubt. M and MET can be instances of two different attitudes and yet nonetheless be parts of a single conscious vehicle or brain state. I am (unconsciously) thinking that I am doubting.

It may indeed be true that we do not normally *say* things like "John believes (or thinks affirmatively) and doubts that it is raining." Now it might seem that thinking and doubting are in opposition to each other in the sense that a person is both affirming and doubting that it is raining. But this is not the structure of the conscious state according to the WIV. It is not that a subject is both believing *and* doubting that it is raining. Rather, the subject is unconsciously thinking (affirming) that she is consciously doubting that it is raining. To use another example, one might also unconsciously think that one wants to eat some pizza (and thus have a conscious desire). But in each case, the content of each attitude differs: one is directed at the world, and one is directed at the mental state. If one were asked about such a situation and then had the corresponding introspective state ("consciously thinking that I doubt it is raining"), then one would be consciously thinking about the desire or belief. It is actually quite odd for a HOT theorist to rely so heavily on "folk psychological" explanations here. Adherence to the Transitivity Principle is indeed intuitively appealing for reasons we saw in chapter 2. However, one also rarely, if ever, says things like "I am (unconsciously) thinking about my desire for pizza." But this is precisely what the standard HOT theorist claims is really going on when one has such a first-order conscious desire.

It may be that part of the reason to resist the notion that two attitudes can be represented in the same vehicle is due to the (mistaken) background assumption that the attitudes in question are both *conscious*. That is, how can someone have or hold in mind such a complex state? How could two mental attitudes be represented in the same conscious mental act? But, of course, according to the WIV, one of the two mental attitudes will always be *un*conscious. So this is not a problem.

Following some of the literature in moral psychology (Little 1997), Kriegel (2003c, 486–488) also makes the point that a subject's conscious state containing two attitudes would still be related differently to its two contents. Echoing the points made in the previous paragraphs, the argument has already been made that there are two different "directions of fit" in such cases, and thus there is nothing contradictory or peculiar here. For example, beliefs have a mind-to-world direction of fit, whereas desires have a world-to-mind direction of fit. A belief is meant to match the world, whereas a desire tries to change it in some way. In the cases described earlier, there is no propositional *content* with a double direction of fit. I consciously want some pizza with its world-to-mind direction of fit, but I also assertorically think that I have that desire with its own direction of fit. Indeed, a MET will actually have a mind-*to-mind* direction of fit because of its metapsychological aspect. But in any case, I am *not* both desiring to eat pizza in the future *and* affirmatively thinking that I am eating pizza. Note, however, that the claim is not merely that a desire can *entail* a belief or thought but that two attitudes *are* anchored in a single vehicle. Thus I find little reason to cling to the view that each conscious state (or vehicle) must have only one attitude regardless of the benefits of the background theory.

4.5.3 A "Sum" or "Complex" Account?

This leads to another closely related question: what exactly is the real *ontological* difference between the WIV and HOT theory? Some have claimed that the difference is merely terminological or verbal.[25] Indeed, Rosenthal himself once said that there is no nonarbitrary way of choosing between these two ways of describing higher-order theory (1986, 345).

There really are two issues here. First, on the neural level, Kriegel has rightly argued that whether there are two states or one state is not at all arbitrary (2003c, 488–494). Citing the familiar "binding problem," he explains how the difference may simply depend on whether or not two neural events, N1 and N2, taking place in different parts of the brain, "are synchronized or not. If they are, then N1 and N2 are bound into a single brain state; if they are not, N1 and N2 constitute two separate brain states" (493). I see no reason why a proponent of the WIV cannot agree with this much, at least as a partial explanation of the real (neural) ontological difference between the WIV and HOT theory (though I prefer the feedback loop account).

Second, there is the issue of just what the nature of such "compound" states are. Kriegel (2006) usefully distinguishes between two kinds of "wholes": what he calls a "sum" and a "complex" (cf. P. Simons 1987, chap.

9). This division aligns with Kriegel's earlier (2005) distinction between the mere *compresence* of two mental states and the *integration* of mental states. The basic difference is that a mere *sum* says that what makes two states part of a single mental state is merely our decision to treat them as such. Once again, this is a purely verbal or stipulative difference that Kriegel (2005) calls the "conceptual-relation strategy." In contrast, a *complex* is a sum whose parts are essentially connected, or bound, in a certain way. A *psychologically real relation or integration* exists between the parts. Thus Kriegel calls it the "real-relation strategy" (cf. Kriegel 2009a, 222). For a sum to go out of existence, one of its parts goes out of existence, whereas a complex can go out of existence even if none of its parts goes out of existence but because the interrelatedness of the parts is altered or destroyed. For example, the state of Hawaii is not merely a sum of seven islands, that is, not merely the geographic combination of islands. It is a complex of islands because it also involves a particular and essential political or governmental interconnectedness (Kriegel 2009a, 221).

Kriegel (2005, 2006, 2009a) has sometimes construed the WIV as a sum account, whereas he understands his own "cross-order information integration" (or "same-order monitoring") model and Van Gulick's Higher-Order Global States model as complex accounts. I strongly disagree with this characterization of the WIV, though it is perhaps understandable how one might initially take it that way, particularly given the admittedly embryonic form the theory took in my 1996 book. However, I do think there is strong evidence to indicate that my WIV is a complex account, even going back to Gennaro 1996. One purpose of this section has been to make this even clearer.

First, in criticizing Rosenthal's HOT theory, I spoke of how the "very *nature* of conscious states is colored by the concepts [in the METs that are] brought to bear on them" (Gennaro 1996, 29). I urged that the MET actually changes the nature of the conscious state, so that, unlike HOT theory, the object of a MET is not merely passively there unaltered by the MET.[26] Second, I had already criticized Rosenthal's belief in unconscious *qualitative* states because the conceptual activity in the METs is essential to the very *identity* of the overall conscious state it is part of. So "a nonconscious qualitative state, *contra* Rosenthal, could not be the *very same state* as the conscious one because of the lack of conceptualization" (Gennaro 1996, 30). That is, when M becomes conscious as part of a CMS, it is not just the same state with consciousness added to it. Third, as was mentioned earlier, I had already elaborated on and emphasized the Kantian-style thesis that it takes the appropriate cooperation between the "sensibility" and the

"understanding" to produce the resulting conscious state (Gennaro 1996, chap. 3).[27] The WIV thus embodies a real-relation strategy and is a complex, not sum, account of state consciousness. If the interconnectedness between M and MET is absent, then there will be no resulting conscious state even if the parts remain intact.[28]

In light of this and the discussion in the previous subsection, it is worth noting that, in a striking passage, Rosenthal himself says that "on the HOT hypothesis, *a conscious state is a compound state*, consisting of the state one is conscious of [i.e., M] together with a HOT. So the causal role a conscious state plays is actually the interaction of the two causal roles. . . . This explains how a state's being conscious may to some extent matter to its causal role" (2002a, 416; italics mine). This comes in response to Dretske's (1995, 117) charge that Rosenthal's HOT theory is unable to explain how a mental state's being conscious could have any function; that is, it would seem that the state's being conscious would make no difference to its causal role. Thus Rosenthal's HOT theory seems to threaten to make consciousness merely epiphenomenal (i.e., without any causal efficacy) because it construes the HOT as a distinct extrinsic state to its target state. Moreover, Rosenthal himself invites this interpretation, since he has certainly sometimes very much minimized the causal role of a state's being conscious (such as in Rosenthal 2002a, 416–417). More recently, however, he explicitly states that his "conclusion about function . . . does not imply epiphenomenalism," although he does also argue that "the consciousness of intentional states has no significant function" (Rosenthal 2008, 831). As we will see in chapter 8, the same kind of problem haunts Carruthers's view.

Although Rosenthal is reluctant to concede the point, he does seem to agree that the WIV and HOT theory differ in at least one important way, namely, that something more like the WIV can better explain how the causal/functional role of a single conscious state can be importantly different from the relevant target state stripped of its HOT. He is rightly suggesting that (at least sometimes) M and MET, when combined, can form a uniquely new state, at least in terms of its functional role, which is therefore not a mere "sum." This sounds more like the WIV than standard HOT theory, particularly given Rosenthal's note that the "interaction of the two roles may not be [merely] additive" (2002a, 421n48). Moreover, he thus seems to be able to make at least some sense of the one-state, two-attitude view. Of course, Rosenthal could simply respond by holding that since a HOT is also a mental state, it too has causal efficacy in addition to the causal role played by M. Thus M and a HOT *together* will have a different causal role from M without a HOT. The main concern I have, however, is that since

M is the conscious state, it is unclear what the causal role of an extrinsic HOT would be. I will not press this issue further.[29]

4.5.4 More on Parts and Wholes

Following up on the theme in the previous subsection, let us look more closely at the structure of conscious states via the notion of mereology, which is the theory of parthood relations, that is, the relations of part to whole and the relations of part to part within a whole (Varzi 2010). Let me be more precise with respect to the WIV without becoming too distracted or delving into the metaphysical weeds. A mereological system requires at least one basic relation. The most obvious choice for such a relation is parthood (or "inclusion"), which we can symbolize as *Pxy* and should be read as "x is part of y." So we can say that M and MET are *proper parts* of CMS, which means that they are parts of CMS each of which is not identical with the whole CMS of which they are part. Thus, for example, the MET is not identical with CMS. So we can understand the WIV as follows:

(WIV) A mental state M of a subject S is conscious if and only if S has a suitable (unconscious) MET, directed at M, such that both M and MET are *proper parts of* a complex conscious mental state, CMS.

More technically, the notion that "x is a proper part of y," often written as *PPxy*, holds if *Pxy* is true and *Pyx* is false. Thus:

X is a *proper part* of y = x is part of y *and* y is not part of x.

So, for example, to say that M is a proper part of CMS is to say that M is part of CMS but CMS is not part of M.

Three other basic axioms can be put as follows:

A1. *Reflexive*: Any object is part of itself. This is usually represented as *Pxx*.
A2. *Antisymmetric*: If *Pxy* and *Pyx* both hold, then *x* and *y* refer to the same object. Or, we might say, two distinct things cannot be part of each other.
A3. *Transitive*: If *Pxy* and *Pyz*, then *Pxz*. That is, any part of any part of a thing is itself part of that thing.

For my purposes, two other important relations are worth mentioning:

Overlap: *x* and *y* overlap, written as *Oxy*, if there exists an object *z* such that *Pzx and Pzy*. The parts of *z*, the "overlap" of *x* and *y*, are precisely those objects that are parts of both *x* and *y*.
Underlap: *x* and *y* underlap, written as *Uxy*, if there exists an object *z* such that *x* and *y* are both parts of *z*. So: *x* underlaps *y* = there is a *z* such that *x* is part of *z* and *y* is part of *z*.

Overlap and underlap are, for various reasons, normally understood as reflexive, symmetric, and intransitive. With respect to the WIV, then, we might say that M (*x*) underlaps MET (*y*), since there is a CMS (*z*) such that M is part of CMS and MET is part of CMS. However, there can also still be some overlap between M and MET insofar as a psychologically real relation holds between M and MET. On the neural level, much the same seems reasonable, since, for example, there may be some overlapping parts of feedforward or feedback loops that extend from M to MET or vice versa. If we construe the vehicles of M and MET in such a manner due to their essential integration, then M and MET can overlap in addition to the underlap.

Returning now to Kriegel's terminology, a sum and a complex can be understood as follows:

Sum: If *Uxy* holds, there exists a *z*, called the "sum" of *x* and *y*, such that the objects overlapping *z* are just those objects that overlap either *x* or *y*.

So there exists a CMS, called the sum of its parts x an y, such that the objects overlapping CMS are just those objects that overlap either x or y.

Complex (or "product"): If *Oxy* holds, there exists a *z*, called the "product" of *x* and *y*, such that the parts of *z* are just those objects that are parts of *both x* and *y*. If *Oxy* does not hold, *x* and *y* have no parts in common, and the product of *x* and *y* is undefined.

So my view is that a CMS contains an underlap of two parts (M and MET) which, in turn, can themselves also overlap with each other. Nonetheless, if there is not a psychologically real integration between the parts, then CMS would not exist in the first place. Having said all this, it is not always easy to translate the foregoing notions from classical mereology into talk about neurons and conscious states. Indeed, philosophers working in this area of metaphysics do not address consciousness much at all to my knowledge. Moreover, they are normally more concerned with nonmental objects, such as tables, ships, lumps of clay, statues, subatomic particles, and so on. They are more worried about questions such as "Can two physical objects occupy the same place?", "When do two combined objects result in another object?" and "What makes a physical object the same object over time?" But the physical configuration of the brain is unique and does not fit naturally into this discussion. Although some sense can surely be made of different parts of the brain, it is again important to note that there are neural connections that go from one part to another or *in* the other. Moreover, the notions of representation and intentionality are not normally encountered in the literature on mereology.

4.5.5 The Infallibility Objection

Another objection to the WIV (or similar views) is the charge that it entails that knowledge of one's conscious states is infallible, especially in light of the problem of misrepresentation discussed in section 4.2 (Thomasson 2000, 205–206; Janzen 2008, 96–99). If M and MET cannot really come apart, then doesn't that imply some sort of objectionable infallibility?

This objection once again conflates outer-directed conscious states with allegedly infallible *introspective* knowledge. In the WIV, it *is* possible to separate the higher-order (complex) conscious state from its target mental state in cases of introspection (see fig. 4.1 again). This is as it should be and does indeed allow for the possibility of error and misrepresentation. Thus, for example, I may mistakenly consciously think that I am angry when I am "really" jealous. The WIV properly accommodates the anti-Cartesian view that one can be mistaken about what mental state one is in, at least in the sense that when one introspects a mental state, one may be mistaken about what state one is really in. However, this is very different from holding that the relationship between M and MET *within* an outer-directed CMS is similarly fallible. There is indeed a kind of infallibility between M and MET according to the WIV, but this is not a problem. The impossibility of error in this case is merely within the complex CMS, and not some kind of certainty that holds between one's CMS and the outer object. When I have a conscious perception of a brown tree, I am indeed certain that I am having that perception, that is, I am in that state of mind. But this is much less controversial and certainly does not imply the problematic claim that I am certain that there really is a brown tree outside of me, as standard cases of hallucination and illusion are meant to show. If the normal causal sequence to having such a mental state is altered or disturbed, then misrepresentation and error can certainly creep in between my mind and outer reality. However, even in such cases, philosophers rarely, if ever, doubt that I am having the conscious state itself.

This point is not properly recognized by Janzen (2008, 96–99). Although he seems to concede that the WIV can disarm the problem of misrepresentation on the first-order level, he argues that the problem "bedevils Gennaro's theory at the level of *introspective* [states]" (98) because there is a complex MET distinct from the first-order state in this case. However, I'm puzzled as to why exactly Janzen thinks that this problem would arise again at the introspective level unless he is assuming some kind of Cartesian infallibility.

First, as we saw in section 4.2, the initial serious problem of misrepresentation (such as Levine's) is aimed at what the HOT theorist says about *first-order* states given the "splitting up" of first-order conscious states into two

states (unlike other theories). Second, as we have also made clear, when one introspects, I take it that virtually everyone agrees there is a "gap" between the *introspective* state and its target, which also accounts for the widely held view that there is an appearance/reality difference and fallibility at that level. But this is not a problem at all; rather, it is the way that any HOT theorist can accommodate the anti-Cartesian view that introspection is fallible. Just as one can have a hallucinatory conscious state directed at nonexistent objects in the world, one can have a hallucinatory conscious HOT directed at a nonexistent mental state. But even when one hallucinates that there are pink rats on the wall, there is infallible *appearance* of pink rats on the wall. The CMS still exists. Third, as was discussed in section 4.2, confabulated states are best understood as introspective states that either bring about the existence of a conscious state (Hill's "activation") or mistake one state for another. Thus it is puzzling why Janzen uses confabulated states as his main example of a first-order misrepresentation. Finally, when one is in a confabulatory state, we must remember that there is indeed an undisputable conscious state involved, but here it appears at the higher-order level as a conscious HOT (or MET). Thus, though that conscious MET has no *object*, one still experiences *that* state (the MET) as conscious, much as one's hallucination of pink rats on the wall still involves a conscious, but nonveridical, state. Once again the analogy holds, and there is no problem here for the WIV. There can be targetless *conscious* HOTs just as there can be nonveridical hallucinatory outer-directed conscious states.

It is admirable that Rosenthal so clearly wishes to make room for an appearance/reality distinction with regard to our own mental states. I agree with the notion that our *introspective* states are fallible and may misrepresent our "selves" and our mental states. But this distinction applies at the introspective level, not *within* first-order world-directed conscious states. If there is an inner analogy to an illusory or hallucinatory first-order conscious state directed at an outer object, it must be a *conscious* state (= introspection) directed at a mental state. But then this is not a case of an appearance/reality difference between an *unconscious* HOT (or MET) and a mental state M. This is again why we should reject Rosenthal's endorsement of Levine's option one for misrepresentation cases. A lone unconscious HOT without its target is *not* a case of fallible *introspection*.

4.5.6 Weisberg's Objection

Another objection will serve the purpose of bringing together and expanding on many of the foregoing themes. Weisberg (2008) argues that none of the "intrinsic" alternatives presented by myself, Van Gulick, or Kriegel fare

any better than Rosenthal's HOT theory regarding the problem of misrepresentation. I focus first on Weisberg's objections to the WIV. First, a minor point: Once again, I take my view to be a modified version of HOT theory. But Weisberg treats my view (as well as Van Gulick's and Kriegel's) as an example of a "same-order" theory. Terminology aside, however, Weisberg raises some important points that need to be addressed:

(1) Weisberg questions the value of using feedback loops to help us understand how the WIV might be realized in the brain. I return to this topic in chapter 9, but Weisberg rightly notes that I recognize that feedback loops are ubiquitous in the brain, participating in both conscious and unconscious processes. He then wonders why such loops would have any effect on phenomenal consciousness and can ensure that no misrepresentation can occur. The answer is that having a conscious state, according to the WIV, involves having more than *mere* feedback loops. They must be feedbacks *of the right kind*, that is, involving the proper integration of a MET and a lower-order state. Moreover, it is crucial to emphasize that a conscious state should be thought of more like a *product* of such integration, as Kant might say. Thus, if there is no match between the LO state and the MET, then the conscious state will not be produced in the first place. One needs to think of a first-order conscious state as the *outcome* of such a match.

(2) Weisberg similarly questions how my appeal to concepts and Kantian "synthesis" can help to explain conscious states. It is true that Kant did not have in mind how such a synthesis is actualized in our brains, but that should not prevent us from applying Kantian insights to current neurophysiology or theories of consciousness.[30]

More importantly, Weisberg notes that there appear to be unconscious cases of synthesis, such as in masked priming and subliminal perception, which both "demonstrably involve concept application" (2008, 173). He uses an example of subliminally perceiving a stimulus *as* money to explain how we disambiguate the word "bank" in masked priming experiments. All of this is perfectly reasonable, but the problem is that Weisberg has now forgotten the initial motivation for any HO theory, namely, the Transitivity Principle. As we saw in chapter 2, it is most certainly true that there are also unconscious applications of concepts directed at the world. However, only when the appropriate concepts appear as constituents of HOTs (or METs) does the target state M become conscious. Weisberg's example is a case where a concept ("money") is *outer* directed and so not part of a HOT. Thus the synthesizing referred to in Weisberg's case is the wrong kind of synthesis to produce conscious states. There is no *higher-order* application of concepts to one's lower-order states.

The response to Weisberg's objection can also be used against a similar point made by Levine (2006, 192), who also wonders why misrepresentation cannot occur in a theory more like the WIV. Levine claims that what matters most is *not* neural binding or integration but whether or not M and MET are *psychologically* bound. As I have made clear, however, there is important psychological binding between M and MET through the synthesizing of passive perception and active concept-applying METs.

I suspect that the underlying concern at the heart of Levine's and Weisberg's objections has more to do with the following questions: If first-order misrepresentation is really impossible in the WIV, then why even call the relationship between MET and M a *representational* one? What kind of representation (and thus representational theory) does not really allow for misrepresentation? The main reason to treat the MET as a representation of M is that it is still an unconscious metapsychological thought about M. Now, at the neural level, the MET is still directed at M even though there is important interaction between them, which warrants treating the complex state as a single state with two contents. As is well known, the brain has layers of representation going from "lower" to "higher" areas. We can think of this in terms of a hierarchy where the higher areas represent the lower areas. But in the case of conscious states, the relation between the MET and M is what Feinberg (2000) would call a *nested* one; that is, there is dynamic interaction in both directions due to feedback loops and concept application. This contrasts with, for example, the central nervous system in general, where we have a nonnested hierarchy, that is, a purely bottom-up sequence of representations.

Feinberg (2000, 2001, 2009) has argued for what he calls the "nested hierarchy theory of consciousness" (NHTC). According to Feinberg, in a nonnested hierarchy, lower and higher levels are independent entities in which the top of the hierarchy is not physically composed of the bottom. A nonnested hierarchy has a pyramidal structure with a clear-cut top and bottom with the higher-levels controlling the lower levels, analogous to a military command structure. In a nested hierarchy, however, lower levels of the hierarchy are nested within higher levels to create increasingly complex wholes. This idea is also applicable to many other structures in living organisms, such as individual cells. Unlike an account of neural hierarchy that views the brain as a nonnested hierarchy, the NHTC (like the WIV) would treat some areas of the brain as a nested hierarchy when conscious states occur. The idea is that lower-order features combine in consciousness as *part of* (or nested within) higher-order features. So consciousness is not *narrowly* localizable, but it is also not very strongly global. And conscious states are

thus neurally realized as combinations of lower- and higher-order brain features. Thus we can view a conscious mental state as a complex of two parts that are integrated in a certain way. Like the NHTC, essential reciprocity exists between specific neural structures on the WIV. The structures in question are not merely laid upon one another without neural functioning going in both directions.

Thus my view is not merely what has been called a *hierarchical* theory whereby the farther up one goes in, say, one's visual system, the more consciously aware of a stimulus one becomes (Pollen 1999; Lamme and Roelfsema 2000). It is much more of an *interactive* theory such that "once a stimulus is presented, feedforward signals travel up the visual hierarchy. . . . But this feedforward activity is not enough for consciousness. . . . High-level areas must send feedback signals back to lower-level areas . . . so that neural activity returns in full circle" (Baars and Gage 2010, 173). And perhaps the most crucial point is that part of the reason for this may simply be that "higher areas need to check the signals in early areas and confirm if they are getting the right message" (173). If there is no such confirmation, including perhaps a hypothetical case of misrepresentation between M and MET, then no conscious state occurs. I will elaborate further on these themes in chapters 6 and 9.

It is important that we not fall into the trap of FOR and dual-content theory and use a functional notion that is really doing the work, as we saw in the previous chapter. But I believe that the situation is different with respect to the WIV. The synthesizing activity in the MET is not a functional or dispositional notion. It is the MET taking up what enters into our perceptual apparatus and acting on that information. Nothing like Tye's poisedness or Carruthers's dispositional HOTs is present in the WIV. There is important actual *interaction* between M and MET both neurally and psychologically, but it is interaction that stems from a dynamic relationship between representations. The subject's HO concepts must *recognize* the incoming representations as something or other. I offer a detailed account of concept possession in chapter 6, but surely something like being able to recognize objects or properties is necessary for concept possession and application. Moreover, this is clearly a "psychologically real" relation between M and MET. Recognition should literally be understood as *re*-cognition, that is, higher-order recognition of the lower-order state.[31]

4.5.7 Two Final Objections

We are now in a much better position to respond to Droege's complaint that "on Gennaro's theory, it is unclear whether the whole complex state

is conscious and there is something it is like to be in only part of it, or whether there is something it is like to be in the whole complex state. . . . If the whole, then the theory is circular. . . . If part, then either [M] is what is conscious and we are back to [HOT theory], or the [MET] is conscious and the theory is again circular" (Droege 2003, 40–41).

This is certainly an important request for clarification. The first part of the answer is that, in my view, it would be most precise to say that "the whole complex *state* [= CMS] is conscious, and there is something it is like to be in only part of it [= M]," but there are important ambiguities here. Recall the distinctions between the first-person and third-person perspectives and between a conscious state (or vehicle) and its content. Looking at it from a more third-person point of view, we can see that the entire complex (brain) *state* should be understood as conscious. Of course, there is still something it is like to be in the whole state (CMS) if that merely means that when a subject S is in CMS, S is having a subjective phenomenal experience. However, there is something it is like to be in only part of CMS in the sense that S is only consciously aware of the *content* of M from the first-person point of view. We can and should deny Droege's conclusion that we are then "back to HOT theory" because of what we have seen in the previous sections. Since the WIV is a "complex" account, the very conscious content of M is essentially interwoven with the MET. That is, M would not have the content it has without its relation to the MET or if the MET were a completely distinct state from M. The MET is presupposed in the very nature of M's conscious content and is thus part of the same state as M. Moreover, we have seen many other reasons to distinguish between the WIV and HOT theory. The WIV is also not circular because it still attempts to reduce state consciousness to the interaction between two *unconscious* mental states, which are fused together to bring about a unique complex conscious state.

Second, let us also consider the remark by Jürgen Schröder that it is "doubtful whether [the WIV] does really account for our intuition that consciousness is an intrinsic property of our mental states" mainly because "although consciousness is now intrinsic to the *whole* state [= CMS], it is not intrinsic to the mental state which is a part of the whole [= M]. This is so because the conscious-making thought [= MET] is not a *property* of [M], but a mental state of its own" (Schröder 2001, 35n8). That is, since I have just acknowledged that there is only something it is like to be aware of M's content, then doesn't this undermine the initial motivation to make consciousness intrinsic to state consciousness?

The answer is no, but we must again be precise. First, since the WIV is a complex account, it is not quite right to say that "the conscious-making

thought [= MET] . . . is a *state* of its own." Once again, the MET is part of the same state that M is also part of. Second, being represented by the MET *is* a property of M, at least in the sense that the MET contributes essentially to M's conscious content. Third, Schröder seems mistakenly to equate "consciousness" with "the conscious-making thought [= MET]." As we have seen, the MET is itself not conscious at all. Finally, then, the first-person intuition that "consciousness seems intrinsic to conscious states" should be understood as a combination of two claims, namely, that consciousness seems intrinsic to M's *content* from the subject's first-person point of view, and that consciousness is intrinsic to the *state* that M is part of. When it seems to me from the first-person point of view that consciousness is not extrinsic to, say, my conscious perception of a house, I am reflecting on the commonsense "intuition" that consciousness seems intrinsic to my first-person conscious awareness *of the house* and also to the perceptual *state* that includes that conscious awareness. The WIV does accommodate this intuition.

I conclude, then, that the WIV is a more plausible account of state consciousness than standard HOT theory, though the two are similar in many ways. The WIV can better handle the problem of misrepresentation and the problem of the rock while retaining the important virtues of HOT theory. It can also answer the hard problem of consciousness, and so I think that we have established the Hard Thesis. We can and should acknowledge that some form of self-reference is involved in any conscious mental state. However, conscious mental states are not literally directed back at themselves, nor are they made conscious by entirely distinct HOTs. Conscious mental states are complex states with parts such that an unconscious MET is directed at a world-directed M. On this issue, it only remains necessary to argue against Kriegel's self-representationalism. Ruling out this approach will complete the primary case for the HOT Thesis. I now turn to this task.

5 Against Self-Representationalism

In the last chapter, we saw the rationale for preferring the WIV to standard HOT theory. In doing so, we noted that the structure of conscious states includes an element of self-reference. One might therefore think that this opens the door to accepting what has come to be known as the "self-representational theory of consciousness," championed most forcefully today by Uriah Kriegel.[1] In this chapter, I argue at length against his current view, but also against a close cousin that he had previously endorsed.

The notion that there is a self-referential or self-representational aspect to conscious mental states certainly has a long tradition, going back as far as Aristotle (Caston 2002) and, more recently, Franz Brentano (1874/1973, 153), who famously held that "every mental act . . . includes within it a consciousness of itself. Therefore, every mental act, no matter how simple, has a double object, a primary and secondary object."[2] According to the WIV, first-order or world-directed conscious mental states are complex states such that one part of the state is directed at another part. Thus conscious states are to be individuated widely, and consciousness is intrinsic to them. Although the WIV is similar to Rosenthal's extrinsic higher-order thought (EHOT) theory, as I will call it in this chapter, we have already seen that there are also some key differences.

I first lay out the three views to be considered in section 5.1. In section 5.2, I criticize what we might call Brentano's "pure self-referentialism" (PSR), namely, that a conscious mental state is literally directed back at itself. I argue against PSR and show that the WIV is a more plausible theory of state consciousness. As we shall see, the WIV is located importantly between PSR and EHOT theory. In section 5.3, I take a somewhat different approach against Kriegel's self-representationalism, namely, via an examination of whether or not peripheral self-directed awareness accompanies all conscious states. In section 5.4, I reply to several additional attempts to support Kriegel's view.[3]

5.1 Three Views of State Consciousness

For the sake of additional clarity and frequent comparisons, I use the following notation in this chapter:

M = a world-directed mental state
M* = the metapsychological state directed at M
M** = a third-order state directed at M*

I use the acronym "EHOT" for Rosenthal's theory to emphasize that M* is entirely *extrinsic* to, or distinct from, M.

We thus have three positions with respect to first-order world-directed conscious states:

(EHOT) A mental state M of a subject S is conscious if and only if S has a *distinct* (unconscious) mental state M* (= a HOT) that is an appropriate representation of M.

(WIV) A mental state M of a subject S is conscious if and only if S has a suitable (unconscious) metapsychological thought, M* (= MET), directed at M, such that both M and M* (= MET) are *proper parts of* a complex conscious mental state, CMS.[4]

(PSR) A mental state M of a subject S is conscious if and only if S has a mental state M* that is an appropriate representation of M, and M = M*.[5]

All three views take extremely seriously the intuitive notion that a conscious mental state M is a state that subject S is (noninferentially) aware that S is in. Recall that this is basically the Transitivity Principle discussed in chapter 2. By contrast, one is obviously not aware of one's unconscious mental states. The differences lie in how each theory cashes out the expression "aware that one is in." EHOT theory says that the awareness of M is a distinct (unconscious) state M* (or HOT) directed at M. PSR maintains that M* = M; that is, M is literally directed back at itself. The WIV claims that M* (i.e., the MET) is an unconscious part of a complex conscious mental state (CMS) directed at M (which is also part of CMS). In each case, some notion of self-reference is involved. This is perhaps more clear for PSR and WIV, but even EHOT theory says that what makes M conscious is a kind of self-referential (unconscious) thought, namely, "that I am in M" (see fig. 5.1).

5.2 Against Pure Self-Referentialism

One version of self-representationalism is pure self-referentialism. This is perhaps the view most faithful to Brentano's original position. Kriegel

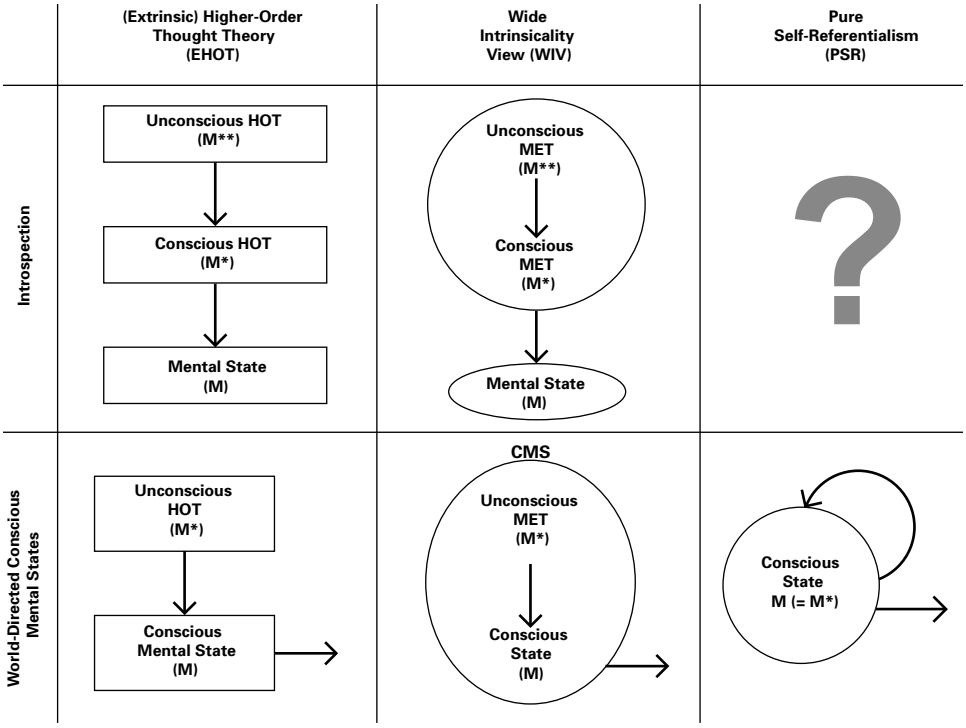


Figure 5.1
Three-way comparison of Rosenthal's "extrinsic" HOT theory, the wide intrinsicality view, and pure self-referentialism.

seems to have argued for it in some of his earlier papers (Kriegel 2003b, 2005). There are a number of interrelated reasons to reject PSR, which, in turn, will also serve as indirect evidence for the WIV.

5.2.1 What Makes a Mental State a Conscious State?

It is fair to say that all three theories try to answer the question "What is the *structure* of a conscious mental state?" However, one real deficiency for PSR is that it does not, to my mind, really attempt to answer the crucial question: what *makes* a mental state a conscious mental state? Both the WIV and EHOT theory are, in part, trying to *explain* how an unconscious mental state becomes a conscious one. Of course, many philosophers and psychologists are not satisfied by the explanation offered, but my point here is that PSR does not really *attempt* to offer much of an explanation at all. Two reasons for this may be that defenders of PSR are inclined to reject

outright all reductive explanations of consciousness (D. Smith 2004, chap. 6; Thomasson 2000, 206) and even to reject the existence of unconscious mental states (Brentano 1874/1973).[6] In this chapter, I argue neither for the existence of unconscious mental states nor for the view that reductive explanations are most desirable (recall sec. 2.2). But to the extent that one agrees with one or both of these assumptions, PSR has a serious problem compared to its rivals.

While both the WIV and EHOT theory answer the question posed in the heading to this subsection with something like "M becomes conscious when an appropriate HOT (or MET) is directed at M," PSR can offer no such explanation. PSR does provide a *description* of the structure of conscious states, but we must distinguish that from the kind of *explanation* we are seeking. If we ask, "What *makes* M conscious?" for PSR, the response cannot be that M* is directed at M because M is supposed to be *identical with* M*. How can M* make M conscious or *explain* M's being conscious if M* = M? Moreover, either M* is itself conscious or it is not, and then the familiar threat of regress (and even circularity) rears its ugly head. If M* is itself conscious, then what makes *it* conscious, and so on? Is the consciousness of M explained in terms of a *conscious* M*? Also, if M* is conscious, then a reductionist account of state consciousness is out of the question. Alternatively, if M* is not conscious, then the PSR defender would first have to acknowledge the existence of unconscious mental states, but even worse, how could M be conscious and M* be unconscious if M = M*?

As I argue later in section 5.2.5, some supporters of PSR may really in the end arguably hold something more like the WIV. Thus, if we also reject EHOT theory, then the WIV represents a superior middle position. It seems necessary to bring in the notion of *parts* of conscious mental states to give an adequate account of state consciousness. In any case, a fundamental question that should be answered by any viable theory of consciousness goes unanswered or is ignored by PSR advocates. To the extent that M* is introduced merely to articulate the structure of a conscious state, PSR may offer a plausible alternative (though obviously I think it is the wrong structure). But if we want M* to play some role in explaining how an otherwise unconscious M becomes conscious, PSR is entirely unhelpful. There is a difference between simply stating that "all (conscious) mental states have a primary and secondary object" and giving an explanation for what makes that mental state conscious.[7]

The PSR theorist might reply that there is a perfectly good explanation for what makes a mental state M conscious, namely, that M becomes conscious when it acquires a particular sort of self-referential content. That is,

M (= M*) has the property of being conscious by virtue of M having the property of being self-referential. However, there are at least two problems with this response, relative to WIV and EHOT theory.

First, the reply still does not help to provide a reductive explanation of state consciousness (which I take to be desirable) because PSR holds that one is conscious (in some sense) of the self-referential content itself. M* is itself conscious in some sense. Recall that all three positions under consideration try to understand state consciousness in terms of some kind of *self-referential intentional* content. Of these three theories, then, it is clear that only PSR cannot offer a reductive explanation in mentalistic terms.

Second, I would again insist that the foregoing explanation is not really *explanatory* but rather *descriptive*. According to PSR, the self-referential content in question is a property of M itself. PSR *describes* the difference between unconscious and conscious states, but M's being conscious is not explained by appealing to M* because M* is identical with M. Now perhaps there is some *other* way for PSR to provide a plausible reductive or, at least, *naturalistic* explanation of state consciousness, but I remain skeptical based on Kriegel's own "argument from physical implausibility" (2003c, 483, 493–496). For example, suppose we understand a naturalistic explanation to include some kind of causal theory of mental content along the lines described in chapter 2. As Kriegel rightly points out, the causal relation is antireflexive, and so we can, at best, make sense of such a relation by invoking talk of one part of a complex state directed at another part. The WIV can clearly accommodate the notion of a causal relation between M and M* combining to produce CMS. Notice, however, that there is still no *pure* self-reference; that is, no conscious (brain) state (or state-part) is literally directed at itself. PSR is indeed physically implausible; it is ruled out if we take the requisite notion of self-reflexiveness too literally.

This problem is made vivid by Buras (2009), who argues that if some mental states are reflexive, then "it spells trouble for causal theories of mental content" (117). That is, if we allow for truly reflexive intentional content as in PSR, then causal theories of mental content must be wrong because causal relations are "irreflexive." But Buras has at most shown that intentional reflexivity is *inconsistent* with causal theories of mental content. Even if we grant this, I suggest that we ought to reject the notion that a mental state can literally be directed at itself. Buras suggests that an alternative strategy would be to allow for some sort of nonrepresentational awareness (or "acquaintance") between M* and M, as opposed to a standard representational relation. He cites the work of Brook and Raymont (2006) and Hellie (2007) in doing so. Such acquaintance relations would presumably

be understood as somehow "closer" than the representational relation and thus can also help to avoid the problem of misrepresentation (and perhaps other problems).

This also seems to be the position of Janzen (2008), who takes Kriegel to task for mischaracterizing some authors in the phenomenological tradition, such as Husserl and Sartre, by insinuating that they would not agree with Kriegel's way of describing the relation between M and M* (Janzen 2008, 116n30). For my own part, I agree here with Kriegel (2009a, 106–113, 205–208) and Buras (2009, 119–121) that this strategy is at best trading one difficult problem for an even deeper puzzle, namely, just how to understand the allegedly intimate and *nonrepresentational* "awareness of" relation between M* and M. I am also inclined to treat it as representational for the reasons given in previous chapters. Finally, it is even more difficult to understand such "acquaintance relations" within the context of a reductionist approach.

5.2.2 Conscious Attention Cannot Be Focused Both Outward and Inward at the Same Time

Another serious problem with PSR is its failure to recognize the implication of the fact that our *conscious* attention is *either* world directed *or* inner directed (but never both at the same time), as even Brentano seems to acknowledge (1874/1973, 128–129). When I am assembling a bookcase or working on this book, my conscious attention is focused outside of me, for example, at the bookcase or my computer screen. If either the WIV or EHOT theory is correct, what makes that state conscious is an unconscious HOT (or MET) directed at M. If, however, I reflect or introspect on my experience, then my conscious attention is focused inward at the mental state itself. But PSR cannot provide such a neat explanation of the difference between outer- and inner-directed consciousness. Leaving aside regress worries, if M* is supposed to be conscious *and* directed back at the *entire* conscious mental state M, then it would seem that M* is directed *both* at the world *and* at one's own mental state M at the same time because, after all, M is supposed to be identical with M*. This doesn't seem possible if it is coherent at all. I certainly may frequently *shift my attention* between, say, the bookcase and my experience of working on it, but I never consciously focus on both at the same time. It therefore also seems that proponents of PSR often slide back and forth between outer-directed consciousness and introspective consciousness without even realizing it.[8]

But if PSR is to explain a world-directed conscious state, M, then it seems committed to the absurdity that M is both directed at the world and at itself

at the same time. Of course, if M* is not itself conscious (as the WIV and EHOT views have it), then PSR either has the problem mentioned in section 5.2.1 (that is, how can M = M* if M is conscious and M* is unconscious?), or leaves us with something closer to the other theories. In any case, PSR supporters do presumably believe that M* is itself conscious in some sense, as we shall see. Unlike the WIV, there is no explicit belief in (unconscious) "parts" of conscious mental states. But the WIV has the advantage of holding that M is an outer-directed conscious part of a complex conscious state (CMS) within which an (unconscious) M* is directed at M. Bringing in parts of conscious states seems unavoidable if one wants to preserve some kind of self-reference in state consciousness.

Perrett (2003) and Zahavi (1998) raise somewhat related objections to PSR. For example, Perrett argues that "there is an inconsistency in Brentano's account. On the one hand, he holds that the content of an awareness is always a proper part of that awareness, where a *proper part* is a part that is not identical with the whole of which it is a part. On the other hand, the secondary awareness is also supposed to possess a content which is identical with itself, since it is its own object. Thus the content of such an awareness cannot be a proper part of itself" (2003, 231). As I have argued, then, if M = M*, then M* would have to be directed back at M *in its entirety*, that is, directed back at M *and* M*. But if M* is itself conscious, then M* is *both* directed at the world *and* at one's own mental state M at the same time. Similarly, Zahavi (1998, 139) describes a "disastrous problem" using Brentano's example of hearing a sound or tone: "A [conscious] act which has a tone as its primary object is to be conscious by having itself as its secondary object. But if the latter is really to result in self-awareness, it has to comprise the entire act, and not only the part of it which is conscious of the tone. That is, the secondary object of the perception should not merely be the perception of the tone, but the perception which is aware of both the tone and itself." The WIV has no such problems. An unconscious MET is directed at only *part* of the entire CMS, that is, the M that is consciously directed at the world. Moreover, the MET is therefore not consciously directed at M, avoiding the conflation with introspection.[9]

One might object that I have thus far ignored a crucial distinction between *attentive* (or "focal") consciousness and *inattentive* (or "marginal" or "peripheral") consciousness. Not all conscious "directedness" is attentive, and so perhaps I have mistakenly restricted conscious directedness to that which we are consciously focused on. The idea is that, in figure 5.1, the "back-turning" arrow for PSR represents inattentive (inner-directed) consciousness, whereas the other arrow represents focused (outer-directed)

awareness. If this is right, then my objection has an easy counterreply, namely, that first-order conscious state M is both attentively outer directed and inattentively inner directed. M* is thus conscious in this inattentive sense. I have three brief replies to this for now but will return to the issue in sections 5.2.4 and 5.3.

First, although it is surely true that there are degrees of conscious attention, it seems to me that the clearest examples of "inattentive" consciousness are outer directed, for example, perhaps some of the awareness in one's peripheral visual field while watching a concert or reading a book. But this obviously does not show that any such inattentional consciousness is *self-directed* at the same time when there is outer-directed attentional consciousness. Second, what is the evidence for such self-directed inattentional consciousness? It is based on phenomenological considerations. For now, it suffices to say that I do not find such inattentive "consciousness" in my experience, which should presumably show up in the Nagelian sense if it is based on phenomenological observation. Conscious experience is often so clearly and completely outer directed that I deny we have such marginal self-directed conscious experience when in first-order conscious states. It does not seem to me that I am consciously aware (in any sense) of my own experience when I am, say, consciously attending to a movie or the task of building a bookcase. Third, when PSR theorists claim to find such inattentive consciousness in their experience, a case can be made that they are philosophically "reflecting" on their experience. But then they are *consciously attending* to their experiences, which is really introspective consciousness. Thus we no longer have a phenomenological analysis of *first-order* conscious states.

5.2.3 PSR Does Not Offer an Account of Introspection

It is also curious that, as far as I know, no clear account of introspection has been presented by supporters of PSR. Perhaps this is simply because they are mainly concerned with first-order conscious states. However, I think the problem goes much deeper for several reasons. First, as we have seen, PSR theorists sometimes conflate introspective consciousness with an explanation of first-order conscious states. If M* is itself conscious, then that seems to indicate the presence of an introspective state, not merely a first-order conscious state. Second, some theorists who otherwise oppose any form of HOT theory are sympathetic to it as an account of introspection. As we saw in chapter 2, Rosenthal reasons as follows: "If a state isn't conscious [at all], there is no HOT. That suggests that a state's being conscious in the . . . [world-directed] nonintrospective way results from something in between

these two [i.e., a nonconscious HOT or MET]" (2004, 24). Both EHOT theory and the WIV can accommodate this important aspect of a theory of consciousness. On the other hand, PSR offers no explanation of just how the transition from first-order conscious states to introspective states might occur. Third, if I am right thus far, it is simply difficult to understand what the structure of introspective consciousness could be according to PSR. If no unconscious thought becomes conscious during such a transition, then does an entirely new state, M**, emerge as directed at M (and therefore also at M*)? Is M** itself also conscious (on pain of regress)? Would M** also be "directed back at itself," so that we would then also have an M*** directed at M**?[10]

5.2.4 The Phenomenological Argument

Perhaps most importantly, then, PSR supporters might argue that M* is conscious (in some sense) based on phenomenological considerations (Kriegel 2003b; D. Smith 1986). Of course, in Brentano's case, M* would have to be conscious because he did not believe in unconscious mental states. There are, as we have already seen, significant problems with holding that M* is conscious, but I now wish to challenge this view more directly, as it is likely at the root of the foregoing difficulties for PSR.

Focusing first on Kriegel's 2003b paper is instructive (cf. 2003c, 485). We find a distinction between "intransitive self-consciousness (or self-awareness)" and "transitive self-consciousness." He first rightly explains transitive self-consciousness in much the same way that EHOT theory and the WIV speak of introspection: "A transitively self-conscious state is *introspective*, in that the object is always one of the subject's own mental states" (2003b, 105). On the other hand, "an intransitively self-conscious state is ordinarily not introspective, in that usually its object is an external state of affairs." So far, so good. Like the WIV, each first-order conscious state contains a metapsychological component (which is a form of self-consciousness or self-awareness, in my view), but the conscious state is outer directed. Moreover, when the shift to introspection occurs, then there is a *conscious* MET directed at one's own mental state. As Kriegel also makes clear, in such transitively self-conscious cases, there is a *further* intransitive self-consciousness accompanying *that* conscious state.

The key point lies in the fact that Kriegel takes intransitive self-consciousness (or M*) itself to be conscious based on phenomenological observation. As I mentioned in section 5.2.2 in addressing inattentional or peripheral consciousness, I strongly disagree and so am much closer to EHOT theory, at least in this respect. M* is not conscious in any meaningful

sense of the term, including the Nagelian sense. For one thing, Kriegel uses a number of vague and mysterious characterizations of M* such as "subtle awareness of having M," "implicit awareness of M," "dim self-awareness . . . humming in the background of our stream of consciousness," and "minimal self-awareness" (2003b, 104–105). I have no objection to these expressions as such, but it is still not clear to me that M* is conscious in any phenomenological sense. Rather, I think it is far better to construe M* as unconscious, and so I prefer the WIV and have often spoken of unconscious "metapsychological thought awareness" and "nonreflective self-consciousness" (Gennaro 1996, 2002). It also does not help to speak of M* as "experienced" or as an "experiential state" (Kriegel 2003b, 121), for this begs the question as to whether or not this "awareness" is phenomenologically *consciously* experienced.

Three more specific problems come to mind. First, the examples that Kriegel uses to illustrate the consciousness of M* really cause us to shift (phenomenologically) to introspection. We are asked, for example, to "suppose . . . that you suddenly hear a distant bagpipe. In your auditory experience of the bagpipe you are aware primarily, or *explicitly*, of the bagpipe sound; but you are also *implicitly* aware that this auditory experience of the bagpipe is *your* experience" (2003b, 104). But it seems to me that the very act of performing this mental exercise results in an act of introspection or reflection. That is, Kriegel is asking us, via our imagination, to focus *consciously* on M (the experience of hearing a distant bagpipe) in "considering" his examples. We are really asked *to reflect on* the hypothetical case in question. But how can we pretend to "consider" such a state of mind without shifting our phenomenological attention onto the mental state or experience itself? To the extent that we can really do so, for what it's worth, I think that our consciousness is completely outer directed, for example, when I am absorbed in a taxing chore or taken with a beautiful painting. Such conscious states can still have the structure of the WIV, but there is no conscious notice at all of M* (= MET). In a sense, then, although Kriegel is not fallaciously conflating introspection and first-order conscious states, he is relying on one's reflective response to make the case that M* is conscious. It is crucial to remember that the WIV also holds that there is an implicit self-referential MET as part of the overall conscious state, but this is not to say that the MET is itself conscious. Indeed, if it were, then we would have a case of introspection, not a world-directed conscious state.

Second, if Kriegel's (or any) phenomenological argument is meant to support PSR, then it also fails due to the lessons learned from sections 5.2.1 and 5.2.2. If we cannot simultaneously *consciously* attend to both outer

objects and inner mental states *and M = M**, then how could M* be conscious? The very same state (without any "parts") would then be both outer directed and inner directed, which is impossible. Moreover, if M = M* and M* is itself conscious, it is difficult to understand how the presence of M* can help to explain why M is conscious in the first place, especially in any reductivist sense.

Now, given the Transitivity Principle, Kriegel is right that "there is something artificial in calling a mental state conscious when the subject is *wholly* unaware of its occurrence" (2003b, 106; italics mine). But this leaves open whether or not such awareness is conscious. Kriegel's use of the expression "wholly unaware" suggests both "consciously and unconsciously unaware," but we might instead hold that a state is conscious when the subject is *unconsciously* aware of its occurrence. Kriegel's argument can really only justify the weaker claim that "there is something artificial in calling a mental state conscious when the subject is *not at least unconsciously aware* of its occurrence." But this is precisely one key issue at hand between PSR and the WIV.[11]

Third, the above critical discussion of PSR is particularly important to the extent that we want a *general* theory of state consciousness. That is, although it is certainly true that there are degrees of conscious attention and self-consciousness, it is desirable to offer an explanation of what *all* first-order conscious states have in common. In the WIV, this is the fact that an unconscious MET is directed at M, both of which are parts of a complex CMS. I suggest that such a general account can only be offered if the MET (or M*) is itself unconscious, because we are so often entirely consciously focused on outer things. Moreover, if we are to allow for, say, animal and infant consciousness, the notion that M* is itself conscious seems highly unlikely. Instead, contra Kriegel, it seems better to hold that any genuine case of a conscious M* is really an instance of introspection, and thus any first-order conscious state is only accompanied by an unconscious M*. In short, holding the Animals and Infants Theses would be more difficult for PSR than for either the WIV or EHOT theory.

The importance of all of this can also be seen in David Smith's (2004, chap. 3) more recent account where he retreats from what appeared to be a previous adherence to PSR (D. Smith 1986, 1989). Smith continues to insist that "the formal analysis of inner awareness [M*] . . . is a task for phenomenology" (2004, 80). However, this leads him to abandon his earlier view that *all* first-order conscious states have such inner awareness (109–116). Smith now allows for basic levels of outer-directed consciousness that lack such self-consciousness or inner-awareness, for example,

when "I am unselfconsciously hammering a nail, or driving down the highway, or choosing to hit the tennis ball crosscourt rather than down the line" (109). Thus "On the view now emerging, inner awareness is an integral part of higher levels of consciousness, realized in humans and perhaps other animals, but it is not present in lower levels of consciousness in humans and other animals" (110). To the extent that Smith no longer advocates PSR, then I agree. However, I suggest that what he should really give up is the view that such "inner awareness" [= M*] is phenomenologically revealed. If he had done so, then he would have recognized that *all* outer-directed conscious mental states can still have a WIV-like structure without giving up the belief that inner-awareness (of some kind) is indeed built into the structure of those states. Smith is correct, however, in allowing for conscious mental states not accompanied by a *conscious* M*. But this should lead one to embrace something more like the WIV instead of the belief that there can be levels of consciousness without any inner awareness whatsoever, especially if one is sympathetic to any of the three positions under consideration.

Thus I disagree with Kriegel that the only reason to posit "such [intransitive] self-awareness . . . is on first-personal experiential grounds" (2003b, 121), and it is also not true that "those who insist that they do not find in their experience anything like an awareness of their conscious perceptions and thoughts probably deny the very existence of intransitive self-consciousness" (121). In the absence of such alleged phenomenological evidence, it is quite appropriate (as Kriegel does) to demand other theoretical and explanatory advantages to positing such self-awareness. Although by no means conclusive, it seems to me that there is ample reason to posit such unconscious METs (see, e.g., sec. 2.4).

In some ways, then, the phenomenological argument lies at the root of the problems raised for PSR in sections 5.2.1 to 5.2.3. If one believes that M* is conscious, then (a) it is difficult to offer any reductionist explanation of state consciousness, (b) one is more likely to conflate introspection with first-order conscious states, and (c) one is unable to offer an account of introspection. I return to additional phenomenological considerations in section 5.3.

5.2.5 Many PSR Views Are More like the WIV in Disguise

Finally, to the extent that PSR is plausible at all, I think it is better construed as the WIV anyway. A close reading of the literature reveals at least some evidence for this claim. Talk of parts and wholes of conscious states abound even when characterizing Brentano's views, not to mention Kriegel's own

extremely useful analysis of various similar positions (Kriegel 2006). It is doubtful that all such references can easily be explained away as merely metaphorical or as a clearly articulated alternative theory.

For example, consider the following sampling of quotations (all italics mine). "The presentation which accompanies a mental act and refers to it is *part of* the object on which it is directed" (Brentano 1874/1973, 128). This suggests that even for Brentano, M* [= "the presentation . . ."] is really only part of the "entire" conscious state. In describing Brentano's view, Natsoulas similarly explains: "Not only is a conscious mental-occurrence instance presented and directly apprehended . . . but also there is, *as part of its occurrence*, awareness of it as this (a mental-occurrence instance)" (Natsoulas 1993, 117). And Smith once said that inner awareness "must be an occurrent *part of* the given mental event itself" (D. Smith 1989, 81). It is difficult to see how M could be *identical with* M* when reading such passages. We are also told by Smith that Brentano's secondary inner consciousness is "a dependent, inseparable *part of* the given act [of consciousness]."[12]

A noticeable shift in emphasis to parts and wholes appears in the more recent writings of Kriegel, who, as we have seen, did argue for PSR in Kriegel 2003b.[13] In addition to the part–whole language used in Kriegel 2006 and 2009a, he had previously said that "a brain state can be said to represent itself if *one part* of it represents *another* part of it" (2003c, 493; italics in original). This comes in the context of preserving some kind of self-representational view within a naturalistic framework, but it also sounds much more like the WIV than PSR. Moving more clearly away from PSR, Kriegel also said that "the mental state yielded by that integration may not actually represent *itself*. . . . At most, we can say that one part of it represents another part" (2005, 48). Much of the same shift even seems to be taking place *within* Kriegel 2009a; that is, in the early part of that book he is more concerned to distance himself from EHOT theory and to avoid the problem of misrepresentation. However, he later explicitly explains that only an *indirect* self-representation is applicable to conscious states (2009a, 215–226). Thus talk of state-parts is much more prominent in the later chapters of Kriegel's book, and it becomes clear that what he calls "direct self-representationalism" is PSR. Kriegel then explicitly says that "there is no *direct* self-representationalism in conscious states" (224). Although the move to indirect self-representationalism is welcome (since it is closer to the WIV), the problem here for Kriegel is that abandoning direct self-representationalism leaves him in a weaker position relative to the problem of misrepresentation. That is, he cannot simply reject the possibility of

misrepresentation on the basis that a conscious state is literally directed back at itself (or its "whole" self).

In the end, then, the case against PSR is very strong. If we are to preserve any useful notion of self-reference at all, something more like the WIV is necessary.

5.3 Another Approach: Peripheral Awareness

Another way to assess the viability of self-representationalism is through a more systematic examination of peripheral awareness in conscious states. We can investigate under what circumstances it is reasonable to hold that peripheral awareness accompanies focal awareness. In some ways, it is a curious fact that two phenomenological claims lie at the heart of contemporary defenses of representationalism. The first phenomenological assertion is that, in addition to our frequent focused (or attentional) awareness of outer objects, we also have peripheral (or inattentional) conscious experience at the "edges" of consciousness. As we have seen, it is sometimes said that some kind of peripheral conscious awareness always accompanies our focal consciousness. Indeed, it seems reasonable to suppose that conscious awareness is broader than those aspects of conscious experience to which one is paying conscious attention. The second claim is the transparency of experience, namely, that when we try to introspect, say, our visual experiences, we "look through them" only to find the outer objects of those experiences. I say that it is a curious fact because many representationalists are often motivated by a desire to reduce consciousness to intentionality without any reference to phenomenal terms. This desire is often accompanied by a decided third-person approach to consciousness and sometimes even a disdain for introspective or phenomenological methods.

In this section, I argue that these two themes are importantly related and can shed light on each other. I lay out four distinct theses on peripheral awareness and show that three of them are true. However, I then show that a fourth thesis, most commonly associated with self-representationalism, is false. Moreover, some of my diagnosis as to why the fourth thesis is false and why the first three are true involves discussion of the transparency of experience. Finally, I respond to several objections and to further attempts to show that thesis four is true. What emerges, once again, is that if one wishes to hold that some form of self-awareness accompanies all outer-directed conscious states (as I do), one is better off holding that such self-awareness is itself *unconscious*, as is held, for example, by standard HOT theory and the WIV.

5.3.1 Varieties of Representationalism

Recall the following flavors of representationalism:

(1) First-order representationalism (FOR)

(2) Higher-order representationalism (HOR)

Recall also that when a conscious mental state is a first-order world-directed state, the HOT or MET is *not* itself conscious. When it is itself conscious, there is a yet higher-order (or third-order) thought directed at the second-order state. In this case, we have *introspection* that involves a conscious HOT or MET directed at an inner mental state. For the sake of this discussion, I mostly ignore the difference between HOT theory and the WIV because they have in common the central claim that the HOT (or MET) is itself normally unconscious.

(3) Self-Representationalism

In this section, I mainly have in mind Kriegel's more recent view (2006, 2009a), that is, the "indirect" self-representational theory, which maintains that the metapsychological state in question is itself (peripherally) conscious and intrinsic to (or part of) the overall conscious state. Thus the idea is that conscious states do represent themselves in some sense, which still involves having a thought about a mental state, just not a distinct or separate state.

5.3.2 Four Theses on Peripheral Awareness

To examine the notion of peripheral awareness in a systematic fashion, it will be useful to recognize from the outset that peripheral awareness could be directed at the outer world or directed back at one's own mental states as some form of peripheral "self-awareness." Moreover, peripheral (or inattentional) awareness is obviously to be contrasted with focal (or attentional) awareness, for which there are again two possibilities: outer directed (perception) or inner directed (introspection). We therefore have four possible combinations: (1) outer focal, outer peripheral; (2) inner focal, inner peripheral; (3) inner focal, outer peripheral; (4) outer focal, inner peripheral. Thus, using the obvious corresponding abbreviations, we have the following four theses:

(OFOP) We (at least sometimes) have outer focal consciousness accompanied by outer peripheral conscious awareness.

(IFIP) We (at least sometimes) have inner focal consciousness accompanied by inner peripheral conscious (self-)awareness.

(IFOP) We (at least sometimes) have inner focal consciousness accompanied by outer peripheral conscious awareness.

(OFIP) We (at least sometimes) have outer focal consciousness accompanied by inner peripheral conscious (self-)awareness.

I propose to examine each thesis separately, at least to the extent possible. Which of the four are true, and which are false? Why? What examples can we use to illustrate each thesis? What, if any, interesting connections exist between them? It is crucial, however, to make two preliminary points: First, it should be clear that each of the four theses is logically independent of the others. That is, none of them logically follows from any of the others, though there may be other analogies and relations between them. Second, we should keep in mind the distinction between conscious attention and conscious awareness. It seems reasonable to suppose that conscious awareness is broader than those aspects of conscious experience to which one is paying conscious attention. We consciously attend to only a subset of that which we are conscious.[14]

(1) *OFOP*: This is certainly the least controversial of the four theses. Nonetheless it is still important to be clear about why. For example, classic examples from visual perception tend to confirm OFOP and the idea that conscious awareness is broader than (focal) conscious attention. I am now consciously focused on my computer screen, but it is not as if everything else in my visual field has "gone dark." I do sometimes shift my attention to the keyboard or the papers and books on my desk, but then the same goes for those experiences. Whatever I may be focusing on in the outer world, I have an accompanying peripheral consciousness of objects in my peripheral visual field.

Importantly, the same seems to go for the other modalities; for example, when I am consciously focused on listening to a Jimmy Page guitar solo in a Led Zeppelin song, it is not as if that's all that occupies my conscious awareness. I am still, in some lesser peripheral sense, consciously aware of the bass and drums in the background. Finally, a case can be made that there are cross-modal instances of OFOP, such as the peripheral tactile sensations and auditory experiences I have when I am (visually) consciously focused on the computer screen.[15] All of this seems to be fairly uncontroversial support for OFOP. While it is true that one can *also* acquire information entirely unconsciously (as in subliminal perception and blindsight), these cases seem quite different from outer peripheral conscious awareness. After all, when I am typing and looking at the computer screen, it is not as if I am like the blindsight patient with respect to the rest of my visual field.

It is also clear that all three representational theories noted in the previous section can and should recognize the truth of OFOP. There does not seem to be any reason for them to reject OFOP, nor does the truth of OFOP conflict with any of the other theses.

(2) *IFIP*: This thesis differs in interesting ways from OFOP. I am inclined to think that it is true as a matter of phenomenology, but it is less clearly so than OFOP. IFIP is basically claiming that, at least sometimes, when I introspect (or consciously think about) my mental states, I am peripherally aware of other mental states that I am not currently focused on. I say "peripherally aware of other *mental states*" because of the "I" in IP. My focus is inner; hence the "IF." However, my peripheral awareness is also inner; hence the "IP."

What would be an example of IFIP? Well, suppose that I am introspecting with the purpose of deciding what I believe about something, say, the death penalty. Provided that we allow for some kind of minimal specious present, it seems phenomenologically accurate to say that I am more aware of some mental states than others during any short interval of introspection. I may be consciously focused on my feelings about the victim's family while, at about that same time, only peripherally aware of my desire to be sure that innocent people aren't put to death. Or I may be consciously focused on my belief in equal justice under the law while only peripherally aware of my sympathy for the victim. It does indeed seem that there is a phenomenological aspect to the peripheral (self-)awareness in these cases. Such mental states seem to be "in the background" even when I am consciously focused elsewhere. Another case might be thinking about my desire for another piece of cake. This conscious thought about my desire can certainly take center stage in my phenomenology. However, it seems to me that I often at the same time (or nearly the same time) have nonfocal (i.e., peripheral) awareness of feeling guilty about having another piece or of the belief that having another piece is not healthy. It seems that when we introspect, we are often able to hold a number of mental states in mind, though some of them will not be objects of focal consciousness.

Notice, however, that we must take care not to construct a case where one's conscious focus simply shifts from one inner state to another, for that would show only that the IF part of IFIP is true, which can easily be demonstrated by any sequence of introspected mental states. Of course, we should again presumably treat the "moment" of introspecting as a short specious present, but we must take care not to let IFIP simply reduce to IF depending on the examples used. We can perhaps talk of a brief "temporal spread" analogous to the spatial spread involved in outer perception, but that still

seems quite different from the considerations that favored OFOP. There is, to be sure, however, something more slippery about IFIP than OFOP.

Perhaps the reason for this is that, unlike a classic visual perception case of OFOP, there is no spatial array of objects when one introspects such that we can mark off a distinction between focal and peripheral consciousness. There is no spatial array of objects to which one can appeal in support of IFIP. Now this observation has its best-known origins in the Kantian view that, while the outer world is revealed to us both spatially and temporally, inner sense has time as its only "form of intuition." That is, we must experience the outer world as spatial and temporal, and thus, according to Kant, space and time are the "forms of outer sense." However, Kant urged that time alone is the form of inner sense. Some have even tried to use this phenomenological fact about how introspection differs from outer experience to argue that the mind–body problem itself can never be solved due to the inherently different perspectives involved (McGinn 1989, 1995). As indicated in chapter 2, I am not convinced that such a drastic "mysterian" conclusion is warranted by these facts. Nonetheless the differences between IFIP and OFOP may be an underlying source of the sense that what is an essentially first-person activity (consciousness) cannot be explained in terms of third-person methods that necessarily involve a spatial component and thus the application of spatial concepts. Of course, we do sometimes say that "such and such is in the back (or corner) of my mind," suggesting a spatial analog to outer perception. But it is not clear that this is anything more than metaphorical language use.

Once again, it seems clear that all three representational theories can and should recognize the truth of IFIP. There does not seem to be any reason for them to reject IFIP, nor does the truth of IFIP necessarily conflict with any of them. However, if something like the "inner perception" or HOP model were true, one might expect the analogy to OFOP to be stronger than it is. As we saw in chapter 3, even HOP theorists do not claim that introspection is perception-like in the same way as outer perception.

(3) *IFOP*: This might seem to be a curious combination. How can we be focused on some (inner) mental state while we are peripherally aware of outer objects? Nonetheless I think that IFOP is true and something like it occurs often. Take the common experience of daydreaming while, say, listening to a lecture. Presumably one is introspecting about something, for example, consciously thinking about one's own mental states. One might think about one's own desire to be with one's children or about one's belief that the lecture is half over. We sometimes "zone out" while watching a television show and realize that we are thinking about something else in a

deep introspective way. Now, in these cases and much like OFOP, it is not as if one becomes blind or deaf to one's outer surroundings. Unlike OFOP, however, IFOP has it that *all* of one's outer consciousness is peripheral to one's inner focal consciousness. The reason is perhaps very simple, namely, that one's sense organs are still functioning and are consciously able to pick up information coming in from the outside. However, one is not consciously attending to that incoming peripheral information. The same goes, I think, for the common experience of trying to remember something, say, for a test or in response to a question. I am in deep introspective thought directed at my own mental states, such as memories and beliefs. When I turn my attention inward to remember something, it is again not as if I don't also consciously see anything outside of me (unless I close my eyes, of course). However, such outer awareness is peripheral to the main focus of my consciousness, which is inner directed.

Instructive here is the much-discussed case of the long-distance truck driver who has been driving for a long time and suddenly "comes to" realize that he has been driving for a while without being consciously aware of the road (Armstrong 1968, 1981). Despite the temptation to agree with Armstrong and say that such outer-directed perceptual states are not conscious at all, it seems to me that this is better described as another good example of IFOP. Indeed, there seems to be a growing consensus that something like IFOP is the best way to handle this case. Some who are otherwise sympathetic to something like Armstrong's HOP theory seem to support such a view.[16] Leaving aside the issue of whether or not the long-distance truck driver case can be used to support HOP theory, Lycan and Ryder (2003) do not believe that the driver (on autopilot) is entirely unconscious of the road. They state, for example, that we must "distinguish normal attention to the road from merely (or minimally) perceiving the road, because the driver does perceive the road but does not have a normal level of attention to road-features perceived, or possibly to any external-world features at all" (133–134). Similarly, they assert that "the autopilot driver does not completely lack awareness of the road (he does perceive the road features, or he would crash), though it is fair to say that he has only a low degree or minimal type of awareness of it" (134).

The same point is made more forcefully by Wayne Wright (2005), who offers empirical evidence in support of the view that without some form of outer conscious perception, the driver would very likely crash, given the cognitive demands of driving. And more to the point of IFOP, Wright correctly explains that since "the distracted [autopilot] driver manages to keep his car on the road for a considerable time, what we should instead

conclude . . . is that he is the subject of visual states of which he has at least some minimal awareness" (46). Wright notes that there is only so much attention to go around and that attention comes in degrees, so that the "highly distracted driver is having what we might regard as very dim experiences [of the road] that receive greatly reduced attention. . . . My claim is that the distracted driver is subject to visual states that are accompanied by enough awareness, a sufficient amount of attentional resources, to enable him to keep the car on the road" (47). As I noted earlier, perhaps part of the reason for this support of IFOP is simply that, as long as our eyes and ears are open while driving long distances lost in thought, it is still the case that our sense organs are functioning and so are consciously able to pick up information coming in from the outside. In the case of driving long distances, the combination of self-preservation and the cognitive demands of driving seem to dictate that one is at least peripherally consciously aware of the road and one's outer environment.

Now, in other cases, such as when one is reflecting on what one is currently seeing, I believe that IFOP can also be helpfully explained by reference to the transparency of experience. Here we (focally) introspect perceptual states that are, at the same time, (peripherally) directed at their very content. Whatever one thinks of the transparency of experience (Kind 2003), it is surely at least important to note that we are not introspecting the outer objects themselves (Stoljar 2004). Indeed, the expression "transparency of experience" is somewhat of a misnomer; it is really the transparency of *introspection*. At minimum, we must keep in mind the distinction between the introspected *state* (or vehicle) and its *content* (or what it is about). A belief in the transparency of experience, as Wright explains, "does not entail . . . that one is incapable of attending to one's visual states, only that when one attends to one's visual state, all that one will find are the features that figure into the state's content" (2005, 63). Thus, given the transparency of experience, we can see how IFOP can easily occur. I may be introspecting my visual experience of the red tomato, but I become (peripherally) aware of the state's content, that is, the red tomato itself.

The larger issue at hand is typically cast as the main dispute between representationalists and nonrepresentationalists about the existence of qualia or, more specifically, nonrepresentational properties of conscious experience (Block 1996). If there is more to an experience than its representational content (as Block thinks), then representationalism is false. If we can introspect nonrepresentational properties of experience, then, contra the representationalist, the phenomenal character of experience is not *exhausted by* its representational content. The issue turns, as Block puts

it, on whether or not there are "mental properties of experience that don't represent anything," which he calls "mental latex." In contrast, "mental paint" is characterized as the "mental properties of the experience that [for example] represent the redness of the tomato" (1996, 29).

My purpose in this section is not to engage directly with this dispute, though my sympathies do, of course, lie more with the representationalist. However, we must still take care to distinguish between what is doing the representing (the state or vehicle) and that which is represented (the content). And my main point has been that something like the transparency of experience can help us to understand (at least one way) how IFOP can be true. Kind (2003) likens at least one notion of the transparency of experience, by analogy, to looking through a pane of glass. We can look through the glass to an object on the other side, but we could also focus on the pane of glass itself. I would add that even when one is focusing on the pane of glass, one can still be peripherally aware of the object on the other side. This suggests another analogy in line with the paint metaphor: If my wood fence is painted over with a *clear stain*, I can still look "through it" to see the fence. However, I can also focus, with some additional effort, on the stain itself. When I do so, however, I still perceive the wood fence, at least peripherally. Perhaps the transparency of experience resulting from introspection is more like seeing through a "mental stain" in the sense that one normally sees right through it to its represented object, given the intimate connection between a state and its content. One could also adopt a special reflective attitude (or attentional state) such that one primarily focuses on the mental stain itself, but the represented object would still be part of the phenomenology, albeit peripherally. This is closer to what Loar (2003) calls "oblique reflection," which he argues is still compatible with the transparency of experience. Oblique reflection, Loar claims, is a kind of introspecting not properly recognized by representationalists.

In any case, it seems to me that all three representationalist views can acknowledge the truth of IFOP. This is perhaps clearer, however, for FOR and HOR theories. It is also worth recalling that recognizing something like IFOP (by Lycan and Ryder) illustrates that HOP theory is not really supported by the famous long-distance-driver cases.

(4) *OFIP*: Unlike the previous three theses, I believe that OFIP is false. Indeed, it is simply a weaker version of PSR, which was rejected in section 5.2. If OFIP is false, then a much stronger version of OFIP would also obviously be false. This stronger view says that outer focal consciousness is *always* accompanied by inner peripheral conscious (self-)awareness. Instead of using

the acronym "PSR," I will label this stronger thesis "OFIP*" in this context for the sake of comparison. Thus we have:

(OFIP*) Outer focal consciousness is *always* accompanied by inner peripheral conscious (self-)awareness.

And recall:

(OFIP) We (at least sometimes) have outer focal consciousness accompanied by inner peripheral conscious (self-)awareness.

Once again, we must keep in mind that neither OFIP nor OFIP* follows from OFOP, IFIP, or IFOP. I have already argued in section 5.2 that OFIP* is false. But I am inclined to think that both OFIP and OFIP* are false for a number of reasons, and there is no reason to repeat them all here. As we have seen, Kriegel holds the stronger OFIP* thesis and takes such inner peripheral conscious self-awareness (IP) to be *essential* for one to have a conscious mental state. However, I also disagree with the weaker OFIP for several reasons.

First, recall that although it is true that there are degrees of conscious attention, the clearest examples of inattentive (or peripheral) consciousness are outer directed, for example, perhaps some of the awareness in one's peripheral visual field while watching a concert or working on one's computer. Indeed, these are frequently the kinds of examples used, by analogy, to support OFIP. But cases of OFOP obviously do not show that any such peripheral consciousness is *self-directed* at the same time there is outer-directed attentional consciousness. This is again just to say that OFIP does not follow from OFOP.

Second, what is the most direct evidence for such self-directed inattentional consciousness? That is, what is the evidence for the IP in the OFIP thesis? It is based on phenomenological considerations. I confess that I do not ever find such inner-directed peripheral consciousness (= IP) *alongside* my outer-directed attentive experience. Except when I am introspecting, conscious experience is so completely outer directed that I deny we have such peripheral self-directed consciousness when in first-order conscious states. It does not seem to me that I am consciously aware (in any sense) of my own experience when I am, say, consciously attending to a play or the task of building a bookcase.[17]

Recall from chapter 1 that Kriegel distinguishes what he calls "qualitative character" from "subjective character" under the larger umbrella of "phenomenal character." He explains that "a phenomenally conscious state's qualitative character is what makes it the phenomenally conscious

state it is, while its subjective character is what makes it a phenomenally conscious state at all" (Kriegel 2009a, 1). In his view, then, the phenomenally conscious experience of the blue sky should be divided into two components: (1) its qualitative character, which is the "bluish" component of the experience (or the *what* of the experience), and (2) its subjective character, which is what he calls the "for-me" component (or what determines *that* it is conscious). In short, I have argued throughout this chapter that Kriegel is mistaken in thinking that an experience's subjective character is itself phenomenally conscious.

Third, in an attempt to diagnose the phenomenological error committed by supporters of OFIP, I suggested that they are really "reflecting" on the experiences themselves in such cases (such as in Kriegel's initial bagpipe scenario). If so, then they are *consciously attending* to their experiences, which is really *introspective* consciousness. Thus we no longer have a phenomenological analysis of *first-order* conscious states; that is, there is no longer any OF but instead a shift to IF. And a shift to either IF thesis cannot show that OFIP is true.

Finally, if we are to allow for, say, animal and infant consciousness, the idea that the metathought (or IP) is itself conscious seems highly unlikely. Recall also the discussion of Smith's retreat from what appeared to be his previous adherence to OFIP*****.

Two final points: (1) One might still wonder: Why should *only* OFIP be false? One answer lies in a clear disanalogy between OFIP and IFOP. No analogy holds between IFOP and OFIP primarily because there is no analogue to an open "sensory channel" or "sense organ" for OFIP. When we have a case of IFOP (as we saw), we can make sense of the OP precisely because of open (outer-directed) sensory channels even when we are introspecting. This allows for the somewhat unusual combination at hand. However, nothing like this is the case for any alleged example of OFIP; that is, when we have outer focal awareness, no open sensory channel can justify the presence of IP. Moreover, in cases of OFOP and IFIP, both kinds of awareness are either both outer directed or both inner directed.

(2) As in our discussion of IFOP, the transparency of experience can be instructive here. It can help to explain why we don't *notice* the shift to introspection and mistakenly suppose that we have a case of OFIP instead of IFOP. Since the object of one's introspection (the mental state) "contains" its content, it is easy to shift from IFOP to OFIP. That is, when we "look right through" the mental state to its content during introspection, we are tempted to think that our focal consciousness is really outer and not inner, but this is merely an understandable error given the transparency of experience.

I think we have good reasons to accept the first three theses, but neither OFIP nor OFIP*. So again if one wishes to hold that some form of self-awareness accompanies all outer-directed conscious states, one is better off holding that such self-awareness is itself unconscious, as do both the WIV and standard HOT theory. Finally, it seems to me that advocates of FOR and HOR theories can and should reject OFIP and OFIP*, since these theses are so closely aligned with the self-representational view. The argument in this chapter is also why I claimed in chapter 1 that Kriegel is mistaken that what he calls the "subjective character" (as opposed to the "qualitative character") of conscious states is itself conscious.

5.4 Three More Attempts and a Counterargument

5.4.1 Functional and Evolutionary Considerations

In a separate paper, Kriegel (2004b) offers a somewhat different approach in favor of OFIP or perhaps even OFIP*. He hypothesizes that evolutionary and functional considerations might give us good reason to accept OFIP. Kriegel first, however, plausibly explains how having the OP in OFOP serves a crucial functional role. Given our limited attentional resources and to avoid the risk of informational overload, "the functional role of peripheral awareness is to give the subject 'leads' as to how to obtain more detailed information about any of the peripheral stimuli, without encumbering the system overmuch" (2004b, 181). Although admittedly speculative, Kriegel then extends this logic to OFIP by similarly suggesting that peripheral awareness of mental state M makes it possible for the subject "to easily (i.e., quickly and effortlessly) obtain fuller information about M" should the need arise (181). Since such peripheral awareness is a good thing to have, it is thus not surprising that it would appear in the course of evolution.

We can identify several problems with this attempt to support OFIP. First, once again, it is not clear that the move from outer peripheral awareness (OP) to peripheral self-awareness (IP) is warranted. As we saw earlier, shifting from arguments about OP to IP involves much more than just a "simple extension" of reasoning. Second, if we are to take this argument seriously and really view it through the lens of evolution, it would seem that we should also extend Kriegel's reasoning about IP to many other animals. However, as we have seen, holding OFIP (let alone OFIP*) makes it that much more difficult to believe in animal (and infant) consciousness in the first place. But if IP is such a good thing to have from an evolutionary standpoint, then it would seem that many other animals should also have it.

Third, it is unclear why we are forced into believing that we and other animals have IP based on this approach instead of, say, having *unconscious* HOTs or METs (during outer-directed focal consciousness). Indeed, there are perhaps equally good evolutionary reasons to suppose that actual unconscious METs accompany outer-directed consciousness. Unconscious METs can presumably become conscious more quickly, resulting in *introspective* conscious mental states. The ability of an organism to shift quickly between outer- and inner-directed conscious states is surely a crucial practical and adaptive factor in the evolution of species. For example, an animal that is able to shift back and forth between perceiving other animals (say, for potential food or danger) and introspecting its own mental states (say, a desire to eat or a fear of one's life) would be capable of a kind of practical intelligence that would be lacking otherwise. Having unconscious METs can thus be understood, from an evolutionary perspective, as a key stepping-stone to the capacity for introspection. In a sense, then, I agree with Kriegel that having *some kind* of self-awareness of conscious states is extremely useful for evolutionary reasons. However, it is unclear why having unconscious self-awareness does not suffice to do the job, or why having conscious peripheral self-awareness will be so much quicker or more effortless in making the shift in question. The kind of information Kriegel has in mind can clearly also be gathered unconsciously and can, in turn, be available to the subject upon introspection. Recall also that supporters of OFIP are sometimes arguably in danger of conflating introspective consciousness with conscious first-order states. In addition, even if such information never becomes conscious at all, it can still be used to guide behavior in importantly relevant ways. After all, it is widely held that unconscious mental states can cause behavior and fill a functional role within an organism.

5.4.2 The Argument from Psychopathologies

Yet another attempt to defend OFIP is made by Ford and Smith (2006). They begin with standard cases of OFOP but then go on to argue that evidence from three psychopathological cases shows how "same-act inner awareness can be an essential feature of every normal contemporary human conscious mental state" (361). Thus they argue for something like OFIP (perhaps even OFIP*), at least for the normal experience of humans. Ford and Smith assert that "within my peripheral awareness there are also—less palpable, as it were—presentations of my own experience, of my passing stream of thought and perception and emotion" (360). Thus they hold that the same conscious state can represent both the outer world and itself. It is an interesting method based on the idea that empirical evidence from

abnormal cases, where some form of self-awareness is missing, can inform us of the presence of self-awareness in the normal case. Without going into great detail, the three cases involve loss of proprioception, amnesia, and depersonalization. Ford and Smith rightly explain just how devastating these psychopathologies can be. Loss of proprioception leads to a serious deficit of "body-image," that is, not being able to feel one's body and a debilitating lack of body control and coordination. Severe amnesia causes the well-known problems of being locked into the present moment and a troubling lack of personal continuity. Depersonalization leads to bizarre cases such that the perceptions and actions of one's body are believed to be happening to someone else. But the key idea, for Ford and Smith, is that if "we find that the impairment or removal of some part of the self-image has an impact on that person's experience, then we may conclude that it must have been present in that person's consciousness, even if the person was not explicitly aware of it" (367).

I am not convinced by their argument for several reasons, though I do find their strategy intriguing and useful. First, as a simple matter of logic, just because the removal of something—for example, normal proprioception—causes deficits in one's conscious mental states, it surely does *not* follow that the *awareness* of that thing is part of normal conscious experience. The relation could be causal instead of constitutive. That is, the typical abilities and awareness in question might merely, in the normal case, *causally contribute* to the phenomenology of one's conscious mental states without being part of the conscious state itself, even peripherally. There are many ways that normal consciousness can be disturbed or impaired, and surely we shouldn't conclude that every such disturbance shows that the item or ability in question normally shows up in our phenomenology.

Second, following on the foregoing theme, it may be that the meta-awareness in question is an *unconscious* awareness of the conscious state. Indeed, this is closer to my own view of such cases, as I have argued at length with regard to severe amnesia (Gennaro 1996, chap. 9). In this way, my view is again closer to standard HOT theory, which might take the meta-awareness (of, say, one's past) as implicit but unconscious rather than a peripherally conscious part of a conscious state. Thus I agree that having such meta-awareness is intimately intertwined with normal conscious experience. Perhaps having implicit episodic memory is even necessary for consciousness at all. I also agree that removing or damaging such meta-awareness will severely impact one's conscious experience. But these facts can equally be explained on my view because, as I noted earlier, lacking such states or having abnormal variants on them can also dramatically

affect the nature of one's conscious states. This can be the case, however, without the I-thoughts or self-image thoughts manifesting themselves in our normal phenomenology. Ford and Smith do not rule out this alternative explanation.

Third, and perhaps most serious, it is not even clear in the first place that any of these examples really involve the awareness of something *mental* at all. That is, if one is going to defend self-representationalism (or even just the more modest OFIP), it must be that the peripheral awareness in question is directed back at the *mental state*. However, Smith and Ford seem to have other intentional objects in mind while describing their examples. For example, in the case of loss of proprioception, what is lost is the sense of one's own body position. But even if proprioception is a peripheral part of all normal conscious outer-directed experience, we are surely no longer talking about OFIP because the IP is directed not back at a mental state but rather at one's own body. To be sure, this could be construed as a kind of self-representation or "bodily self-consciousness," but it is not the kind that supports OFIP. Indeed, this case is arguably closer to an instance of OFOP, since our proprioceptive sense is often directed at the items in the world, such as the tactile consciousness of my pedaling the bicycle while I am riding through town (to use one of their examples). At best, I think we should describe these examples as illustrating that there is an abnormal (but unconscious) I-thought presupposed in the resulting abnormal conscious state. Another way to put it is that the *reference* of the "I" in the I-thought can be one's body without the I-thought itself being conscious at all. Cases of depersonalization can be described in a somewhat analogous fashion; namely, the reference of the "I" in the I-thought is in error, which, in turn, leads to serious deficits of consciousness.[18] In any case, Ford and Smith have not really shown that the features of the self-image to which they refer "must appear as part of each normal conscious experience" (370).

5.4.3 Another Example of OFIP?

Another response in support of OFIP might go as follows:[19] Consider a case where a person, P, is introspecting about something, say, P's philosophical beliefs and thoughts about the meaning of life. We can imagine P reflecting in such a way while sitting in a public park. During this process, a person walks by wearing a strikingly attractive shirt, causing P to shift his attention to the shirt. That is, something outer has caught P's eye and has led P to have outer focal conscious awareness. Given that P had previously been focused on inner states, isn't it reasonable to say here that P has outer focal conscious awareness of the shirt while also having peripheral consciousness

of P's thoughts? The idea is that the inner awareness has (perhaps temporarily) receded into the background of P's consciousness but has not disappeared altogether. This would seem to be a case of OFIP.[20]

My response here is twofold: First, at this point, I am simply inclined to insist that all of P's consciousness is outer focused *at the time* in question. From a phenomenological point of view, if I were in that situation and became so diverted in attention to an outer object, it is not clear to me that the previous focal awareness on my thoughts has merely been pushed into the background of my consciousness rather than disappearing altogether. I fail to see the motivation and evidence for the view that there is *conscious* IP, as opposed to the view that such reflective thoughts became utterly unconscious. Notice also that the case in question stipulates that the outer awareness becomes focal when I see the shirt. Otherwise we would at most have an instance of IFOP and not OFIP.

Second, it is also interesting to note that this alleged example of OFIP importantly differs from the cases considered earlier. In the prior examples, we had been considering standard phenomenological cases where the content of the outer-directed consciousness *matches* the content of the inner peripheral awareness. My outer-directed consciousness of, say, building a bookcase is supposed to be accompanied by inner peripheral consciousness of my building the bookcase. Indeed, this is most certainly what defenders of OFIP and OFIP* have in mind, and it clearly lies at the heart of Kriegel's self-representational account. However, the case described here is quite different, even if one still wishes to use it in support of OFIP. In this scenario, the content of outer consciousness (the shirt) differs from the content of the alleged inner peripheral awareness (my thoughts about the meaning of life). There is no match between the two contents. Thus we could distinguish between the following:

(OFIP-MIXED) We (at least sometimes) have outer focal consciousness *of x* accompanied by inner peripheral conscious (self-)awareness *of a distinct state with content y*.

(OFIP-MATCH) We (at least sometimes) have outer focal consciousness *of x* accompanied by inner peripheral conscious (self-)awareness *of the awareness of x*.

As I noted earlier, OFIP-MATCH is the standard self-representationalist reading of OFIP, which I have argued is false. I suppose it is possible to accept OFIP-MIXED and not OFIP-MATCH, although I do not think that such a move is warranted or well motivated. Even if a stronger case could be made for OFIP-MIXED, it is clear that OFIP-MATCH would not follow.

Still, one might ask: why *can't* OFIP be true? Well, first, it is unclear what the strength of "can't" is. Perhaps it is not *logically* impossible, but OFIP still seems to me actually false. Second, this way of putting the question, I think, unfairly shifts the burden of proof onto the skeptic of OFIP. It seems to me that the supporter of OFIP has the burden of proof; she is making the positive and existential claim about the existence of a certain state of mind. The onus is on the supporters of OFIP and (especially) OFIP* to offer evidence and reasons for their view. Part of what I have done in this chapter, then, is simply to critically examine arguments that have been put forward for OFIP and OFIP*, finding them to be seriously lacking. It should not be up to the critic to show that either claim is somehow impossible.

Finally, it might be urged that I should back off the *strong* claim that IP *never* accompanies OF, which follows from my rejection of OFIP. I am not inclined to do so for many of the reasons already adduced in this chapter, though I am willing to concede that David Smith's 2004 hybrid position is *possibly* correct and that this is not an all-or-nothing issue; that is, maybe *some* OF states are accompanied by IP, but not all. Nonetheless, to the extent that we do treat the matter as all-or-nothing, I am much more inclined to endorse the strong claim that OFIP is false. This is primarily because I am so much more certain, from a phenomenological perspective, that there are cases where our conscious attention is entirely outer directed than I am sure that there are some cases where IP accompanies OF. In other words, instances of OF without IP are so much clearer to me than the extremely elusive and slippery notion that IP (at least sometimes) accompanies OF.

5.4.4 A Kriegel Counterargument

Kriegel (2009b, esp. 377–379) has responded to some of the arguments presented here (and in Gennaro 2008a). I respond to three of his points:

(1) Perhaps his most important and interesting line of argument is that there may be a very good reason why *introspection* does *not* reveal peripheral inner awareness. According to Kriegel, this is because when one introspects one's current conscious experience, that inner awareness "displaces" or "annihilates" any peripheral inner awareness that was present. The previously existing peripheral inner awareness is transformed into a focal inner awareness. Part of Kriegel's motivation is to show that the transparency of experience is compatible with OFIP*, as well as to explain why we cannot "catch ourselves" in a state of inner peripheral awareness.

Several points need to be made here. First, I largely agree with Kriegel's discussion along these lines, namely, the idea that *introspection* could not reveal any alleged *peripheral* inner awareness (accompanying outer focal

awareness) because it supplants or annihilates it. I think he is right that any argument that only attempts to show the falsity of self-representationalism (or OFIP*) based on the transparency of experience is misguided. However, it is important to recognize that this point does not show that any IP accompanies OF in the first place. I have already cast significant doubt on the claim that there is any such IP.

Second, none of my arguments proceeded along the lines above. Rather, I used the transparency of experience as a possible alternative explanation for why someone might *confuse* what is really an *IFOP* for an OFIP. I don't rely on transparency itself (and thus *introspection*) to reject OFIP, as perhaps some others might (Kriegel 2009b, 371–376). When I claim not to find peripheral inner awareness in my phenomenology, I am trying to restrict myself to the nonintrospective cases. Indeed, I think that it is the *supporters* of OFIP who are really introspecting, thus having IFOP and not really OFIP. So, again, I agree that *if* we had a case of OFIP, then transparency cannot refute it. But that does not show that there is OFIP in the first place. At the least, my alternative explanation is an equally plausible one.

Third, it is revealing that Kriegel has now all but conceded that there is *no introspective phenomenological evidence for his position*. This is a rather incredible concession given that self-representationalism is supposedly based on first-person phenomenological considerations. Of course, Kriegel still claims that there is some nonintrospective phenomenological evidence for what he calls the "general impression" that IP always accompanies OF. But, again, this rather mysterious notion has been rejected in this chapter for many reasons. If IP is as ubiquitous as Kriegel says, one would suppose that thinking of such cases would be easy and uncontroversial. Moreover, any such evidence for OFIP would then seem to rely on *memory* as opposed to contemporaneous phenomenological facts. And surely memory is even less reliable than phenomenological facts. I am not unreasonably demanding that there be an entirely uncontroversial case of OFIP, but it is reasonable to examine whether or not we have as good evidence for OFIP as we clearly have for the other cases.

(2) Another important line of argument is Kriegel's repeated insistence that it is somehow so arbitrary or odd to hold that there are cases of OFOP, IFIP, IFOP, but not OFIP (and even OFIP*).

But, first, it is "arbitrary" for me to claim that only OFIP (and OFIP*) is false if I do not give any reasons why the theses differ in key ways. However, that is precisely what I have done. There are several key differences and *disanalogies* between OFIP and the other three theses that make it perfectly

*non*arbitrary to reject only OFIP (and thus OFIP*). For example, I explained why I think that IFIP is plausible whereas OFIP is not. There are also clear examples of OFOP that are lacking for OFIP. I also explain why IFOP makes perfect sense, such as in the introspecting/watching television case, given that there is an open sensory channel directed outward. Thus there is nothing "odd" or "arbitrary" here at all.

Second, as we have already seen, this way of phrasing the issue unfairly shifts the burden of proof onto the skeptic of OFIP. The onus is on the supporter of OFIP and (especially) OFIP* to offer evidence for such a state of mind. We should not simply *infer* that OFIP* is true because the other theses are true.

Third, many of Kriegel's examples or analogies that allegedly support OFIP are not really cases of OFIP, or at least not cases of OFIP-MATCH, which again is supposed to be the central claim made by a self-representationalist. For example, he talks of an OF case of a television screen with an IP of one's anxiety. He also uses an IFIP example of focal awareness of nervousness before giving a presentation and the accompanying peripheral awareness of such nervousness. But neither of these cases support OFIP.

(3) Last, I strongly disagree with Kriegel about the lack of a priori or conceptual evidence for the more standard unconscious HOTs, as in EHOT theory, or unconscious METs for the WIV (see also Kriegel 2009a, 116–124). But as we saw at length in chapter 2 (especially sec. 2.4) and in chapter 4 (sec. 4.4), we have good reason to treat the Transitivity Principle as a conceptual truth. I also provided several other reasons to opt for something like HOT theory. And we also saw that just because one claims to know a proposition P a priori, it does not follow that no empirical considerations can support or refute P.

It is true that direct phenomenological considerations are not what primarily motivate HOT theory, but Kriegel claims that there is no sign of *indirect* phenomenological evidence, either. He does mention, as one possibility, the wine-tasting argument presented by Rosenthal, but Kriegel dismisses HOT theory as no better than other theories of consciousness in explaining the phenomenon. But this piece of evidence is not to be used in isolation. It should be used in conjunction with the Transitivity Principle and other supporting considerations. Moreover, as Kriegel knows, the case for any theory of consciousness depends to some extent on an argument by elimination, narrowing down the plausible options, as I did in chapters 2 and 3. We must also remember that an important motivation for HOT theory is the desire for a reductionist account of state consciousness, a view not shared by most self-representationalists.

In closing, then, I think that PSR, OFIP, and OFIP* are false. Thus Kriegel's self-representational account is also mistaken. I have argued that OFOP, IFIP, and IFOP are true, and that the transparency of experience can shed some light on some of the reasons why. The primary result is that one is better off holding that any self-awareness that accompanies world-directed conscious states is itself *unconscious*, as is the case for EHOT theory or the WIV.

Overall, then, I conclude that the HOT Thesis is very plausible; that is, a version of the HOT theory is true and thus a version of reductive representationalism is true. We are on our way to solving the Consciousness Paradox set forth in the opening chapter, though it also must be shown how the HOT Thesis is consistent with the remaining theses. I now turn to the Conceptualism Thesis.

6 In Defense of Conceptualism

I have defended the HOT and Hard Theses in previous chapters. I now turn to the Conceptualism Thesis. Conceptualism is roughly the view that the content of perceptual experience is fully determined by concepts possessed by the subject. Exactly what conceptualism is, and therefore what nonconceptual content is, is the main topic of section 6.1. It will also be important to address in more detail the nature of concept possession, as well as the distinction between personal and subpersonal level content. I then argue in 6.2 that conceptualism has a natural affinity with HOT theory, and that each can shed light on the other partly via an examination of the phenomena of ambiguous figures and visual agnosia. Then, in sections 6.3 and 6.4, I reply to the two most common phenomenological objections to conceptualism, the richness of experience argument and the fineness of grain argument. The discussion in this chapter will also force us to revisit the problem of misrepresentation, first addressed in chapter 4, as well as the relationship between consciousness and attention. Overall I argue that conceptualism is far more plausible than the alternative.

6.1 What Is Conceptualism?

6.1.1 Definitions and Motivations

Here are a few representative definitions of conceptualism:

(1) "No intentional content, however portentous or mundane, is a content unless it is structured by concepts that the bearer possesses" (Gunther 2003, 1).

(2) "The way a subject represents the world can be fully specified by using concepts she possesses" (Toribio 2007, 446).

(3) "The representational content of a perceptual experience is fully conceptual in the sense that what the experience represents (and how it

represents it) is entirely determined by the conceptual capacities the perceiver brings to bear in her experience" (Chuard 2007, 25).

Thus the central philosophical issue is whether or not one can have contentful conscious experiences of objects (or properties or relations) without having the corresponding concepts. The basic idea is that, just like beliefs and thoughts, perceptual experiences have conceptual content. So conceptualism is the view that all conscious experience is entirely structured by concepts possessed by the subject; that is, perceptual experience is conceptual through and through (McDowell 1994). This position has some affinity to Sellars's "myth of the given" warning against the possibility of unconceptualized experiences "given" in perception (Sellars 1956).

In a somewhat Kantian spirit, we might say that all conscious experience presupposes the application of concepts, or, even stronger, the way that one experiences the world is *entirely* determined by the concepts one possesses. Indeed, Gunther (2003, 1) initially uses Kant's famous slogan that "thoughts without content are empty, intuitions [= sensory experiences] without concepts are blind" to sum up conceptualism (Kant 1781/1965, A51/B75). It is also abundantly clear that McDowell's conceptualism was greatly influenced by Kant (McDowell 1998).[1]

In any case, let us define conceptualism as follows:

(CON) Whenever a subject S has a perceptual experience *e*, the content *c* (of *e*) is fully specifiable in terms of the concepts possessed by S.

I believe that CON is true. One motivation for it stems from the widely held observation that concept acquisition colors and shapes the very conscious experiences that we have. We have already seen this as a primary rationale for HOT theory, for example, via Rosenthal's wine-tasting example. This view seems widely held by many who are otherwise silent on HOT theory and CON. For example, Siegel (2006) argues that acquisition of certain conceptual abilities can make certain kinds of things, such as Russian sentences and pine trees, phenomenally look different. After one learns Russian (or Cyrillic), there is a phenomenal difference between looking at a written page in the language. The increased understanding of a language causes a phenomenal shift in the experience. Much the same seems true once one has learned how to recognize pine trees in a grove containing many different kinds of trees. In a similar vein, it seems prima facie reasonable to suppose that when one has a perceptual experience *e*, one *understands* or *appreciates* the content *c* of *e* in at least *some* way, which, in turn, requires having and applying concepts.

In addition to phenomenological support, there is also some empirical evidence for this view. Goldstone and Hendrickson (2009) review the literature on so-called categorical perception, which is the phenomenon by which the categories [= concepts] possessed by an observer influence the observer's perception, including both auditory and visual stimuli. For example, making perceptual discriminations between objects is increased when those objects belong to different categories. With regard to color, for example, people show better ability to remember which of two colors has just been shown to them when the colors belong to different color categories. Goldstone and Hendrickson argue that linguistic categories facilitate recognition and influence perceptual judgments. Categorical perception shows that people organize the world into categories that, in turn, alter their perception of the world. Thus categorical perception involves the interplay between humans' higher-level conceptual systems and lower-level perceptual systems. Much as we saw in chapter 4, the visual system is controlled by a network of higher-order areas that have top-down feedback connections to the visual system. Although not conclusive support for the stronger claim in CON, these findings are some evidence for a Kantian-style conceptualist model.

Additional support comes from experimental work in the cognitive neuroscience of attention, which demonstrates a critical role for concept-based attentional priming even in ordinary visual contexts (Kastner 2004). Attentional mechanisms that operate in the visual system appear to be controlled by higher-level areas that generate top-down signals that are transmitted via feedback connections to the visual system. Conceptual knowledge shapes perception as early as the lateral geniculate nucleus (LGN), which seems incompatible with claims made by many nonconceptualists.

Another influential motivation for CON is to explain how perceptual experience can provide adequate reasons for empirical beliefs about objects in the world (Brewer 1999, chap. 5; 2005). The concern here is that if perceptual experiences are not fully conceptualized, then they cannot ground the beliefs based on those experiences. Surely a perception of the cup on the table is the basis for the belief that there is a cup on the table. This is also a key aspect of McDowell's (1994) argument that if our perceptual experiences are to provide us with justification (or reasons) for our conceptualized beliefs, then such perceptions must also be conceptualized. Although I agree with this epistemological line of argument, I do not pursue it in this book. I am more concerned with phenomenological and other arguments relating to the nature of perceptual consciousness.

On the other hand, nonconceptual content has been phrased in various ways and, of course, is an essentially contrastive notion. Consider:

(1) "The central idea behind . . . nonconceptual mental content is that some mental states can represent the world even though the bearer of those mental states *need not* possess the concepts required to specify their content" (Bermúdez and Cahen 2010; italics mine).
(2) "Nonconceptualists maintain that there are ways of representing the world that *do not* reflect the concepts a creature possesses" (Toribio 2007, 446; italics mine).
(3) "What makes the content nonconceptual for subject S is simply the fact that S *need not* herself have the relevant concepts and thus need not herself be in a position to form the relevant thought" (Tye 2006a, 10; italics mine).

I will put the thesis of *nonconceptual content* as follows:

(NC) Whenever a subject S has a perceptual experience *e*, the content *c* of *e* is at least partly specifiable in terms of concepts *not necessarily* possessed by S.

Before moving on, three important clarifications are necessary:

(1) NC must be understood as containing correctness conditions, that is, conditions under which the representational content in question accurately represents the world.

(2) NC is actually closer to what might be called *weak* nonconceptualism. If CON is false, then at least *some* perceptual experiences have at least *some* nonconceptual contents. However, one might hold a stronger version of NC that says:

(STRONG-NC) Whenever a subject S has a perceptual experience *e*, the content *c* of *e* is (or can be) *fully* specifiable in terms of concepts *not* possessed by S.

But since I think that the weaker NC is false, then it would follow that a stronger version of NC would also be false. Moreover, since STRONG-NC is so strong, it is also less plausible. I find it difficult to understand the idea that a creature could have genuine conscious perceptual experiences, without possessing *any* concepts with respect to such experiences, not even some very basic or coarse-grained concepts. No doubt some of my puzzlement stems from adherence to HOT theory, but STRONG-NC still seems to be implausible independently of HOT theory.[2]

(3) One will notice that CON and NC are carefully phrased in terms of the *content* of a perceptual state, as opposed to the state (or vehicle) itself.

Some have drawn a distinction between so-called content nonconceptualism and content conceptualism, on the one hand, and state (or vehicle) nonconceptualism and state conceptualism, on the other hand. Bermúdez and Cahen (2010) explain that for "a state to be *state*-nonconceptual is for the organism undergoing that state not to be required to possess the concepts involved in a correct specification of the contents of that state; the state is then *concept-independent*. Conversely, the mental state would be *state*-conceptual if the organism *could not* undergo the state in question without possessing the concepts involved in such a specification of its contents; the state is then *concept-dependent*." Heck (2000) first articulated this distinction based some ambiguities he finds in Evans 1982.[3] However, it seems to me that the content version is the primary and more important view for several reasons:

(a) Heck himself doesn't take the state view seriously: "I suspect that the state view is indefensible—even incoherent, if coupled with the claim that the contents of beliefs are conceptual" (2000, 486n6). Others have also argued that we should focus primarily on the content view and that the state view "does not bear serious scrutiny" (Bermúdez 2007a, 69; Byrne 2005).

(b) It is somewhat puzzling as to what it even means to say that an experiential *state* is composed of concepts. After all, it is the intentional contents that need to be specified to speak meaningfully about concept attributions and concept compositionality.

(c) As we have seen, in virtually any reductionist representational view, the phenomenal content of a conscious state is exhausted by its intentional content. Thus, if a perceptual experience has no nonintentional features, it makes even less sense to separate the content and state views for the purposes of my discussion. In short, concept conceptualism would seem to entail state conceptualism anyway.

In any case, one motivation for holding NC is that perceptual experience seems to outstrip the concepts that one possesses (Tye 1995, 2006b). Part of the issue centers on just how "rich" the content of conscious perceptual experience is. It seems, for example, that we can experience a complex visual scene, such as a landscape, without having all the concepts of the objects or properties experienced. I address this argument in section 6.3.

Other related alleged support for NC has to do with the so-called fineness of grain in our experience. It is often said that conscious perceptual experience is much more fine-grained than the concepts one possesses (Evans 1982; Peacocke 1992; Kelly 2001a, 2001b; Tye 2006b). In other words, it seems that one can experience many objects or properties without having the concept of that specific object or property. For example, it seems that

a subject could experience a novel shade of red without having the corresponding concept and without being able to reidentify that shade on a future occasion. I address this topic in section 6.4.[4]

It is worth mentioning that some of the impetus behind NC goes back to the work of Dretske (1981) and his distinction between *analog* and *digital* representations. If there is a state of affairs that some object *o* has property F, a representation carries the information that *o* is F *in digital form* if it carries no further information about *o*. But if the information is carried *in analog form*, then it carries additional information about *o*. The basic idea is that much more information is carried in analog form than is delivered in digital form. Information is thus lost when an analog representation is transformed into a digital one, such as a belief or thought. While nonconceptualists would agree that the propositional attitudes represent the world in digital form, they claim that perceptual states are better understood as representing the world in the more "rich" analog form.

Finally, it should be noted that we have already independently rejected two theories of consciousness that rely heavily on nonconceptual content. As we saw in chapter 3, Tye's PANIC theory incorporates this notion (the "N" in PANIC). And Carruthers's dual-content theory also explicitly involves analog content. Although Carruthers takes his view to be a kind of HOT theory, he also somewhat surprisingly understands his view to be a form of HOP theory mainly due to his support for nonconceptual content (Carruthers 2004). If I am right, however, then we have yet another reason to reject their theories.

6.1.2 Concept Possession and the Nature of Concepts

Before moving on, it is necessary to be clearer about concepts. We did explore in chapter 2 the issue of how concepts acquire their content mainly via an examination of the causal view. But what exactly is a concept, and what does it mean to possess a concept?

Much has been written over the past few decades about the *nature* of concepts in philosophy, psychology, and cognitive science. Questions such as "What are concepts?" and "What is it to possess a concept?" are central to these fields and notoriously difficult to answer. One major anthology (Margolis and Laurence 1999) and a number of other important works (Peacocke 1992; Fodor 1998; Prinz 2002; Murphy 2002) have contributed greatly to the debate. Unfortunately these issues are rarely addressed alongside a theory of consciousness.[5]

Some of the issues are familiar and long-standing. For example, are concepts mind-independent abstract objects in some Platonic or Fregean sense,

or are they better understood as mental representations, such as constituents of thoughts? If they are mental representations, it is sometimes said that concepts are to thoughts as words are to sentences. Concepts constitute thoughts. Indeed, I have assumed thus far that concepts are mental representations. The main resistance to this view is that different creatures can share or grasp the same concept (Peacocke 1992). If concepts are purely subjective mental representations, then that would seem to be impossible.

Nonetheless it has been pointed out that this concern can be sidestepped by making a somewhat different distinction between concept *types* and concept *tokens* (Sutton 2004; Margolis and Laurence 2007a). It is true that two creatures cannot have the same concept tokens, since they are particulars in a subject's mind. However, two subjects can share the same concept type. That is, there is little reason to suppose that different tokens cannot be of the same type. And there can even be concept types that have never been tokened in a particular mind. In any case, a common default view in cognitive science is that a thought is constituted by concepts. Some will go further and claim that thoughts are based on wordlike mental representations, or a "language of thought" (Fodor 1975). But my main point here is simply to reiterate and make clearer the commonly held view among representationalists that concepts are tokened instances of mental representations, which is also perfectly consistent with HOT theory.

It also seems clear that *possessing* a concept C involves having some kind of *ability* with respect to instances of C. But which ability? The ability to form an image of C's? To display linguistic competence about sentences referring to C's? To behaviorally discriminate instances of C's from non-C's? It appears that all these options suffer from insuperable difficulties. Counterexamples abound for any proposed answer; for example, having the concept ELECTRON or JUSTICE does not seem to involve forming a mental image, and tying concept possession too closely to linguistic competence is highly problematic because some concepts are arguably possessed by nonlinguistic creatures.[6]

Similar long-standing problems arise for a proper theory of the *structure* of concepts. For example, the classical (or definitional) theory of concepts, according to which simpler concepts express necessary and sufficient conditions for falling under any concept C, has largely fallen out of favor due to the obvious difficulty of discovering just what those conditions are in many instances. Prototype theory tells us that a concept C should be analyzed in terms of a set of *typical* features of C's class. It is a probabilistic or statistical view of concepts (Rosch and Mervis 1975) to be contrasted with the classical model. Exemplar theory analyzes a concept C (say, CAT) in terms of

a particular exemplary instance of a cat (E. Smith and Medin 1981). The "theory-theory" treats concepts as structured representations analogous to theoretical terms in science (Carey 1985; Gopnik and Meltzoff 1997; Keil 1989). According to conceptual atomism, or informational atomism (Fodor 1998), most (if not all) lexical concepts have a primitive structure and thus are unstructured symbols.

Well-known criticisms of each abound, and the literature contains many excellent summaries examining the pros and cons of each theory (Margolis and Laurence 1999, 2008; Murphy 2002; Prinz 2002; Earl 2007). There are also hybrid theories, such as Prinz's (2002) "proxytype" theory. I won't review the literature here, but, just to give one example, informational atomism seems to lead to the radical nativist conclusion that most lexical (if not, all) concepts are innate. Suffice it to say that each theory has significant problems, and various hybrid views are currently being developed.

Moreover, given the plethora of problems associated with each of the theories, compelling cases have been made for the view that there is no single correct theory for all concepts (Machery 2005, 2009; Weiskopf 2009). This is called *conceptual pluralism*, or the *heterogeneity hypothesis*, and says that concepts do not form a natural kind at all. Each of the foregoing theories seems to work well for one or two kinds of concepts, but not for all. Ideally, it is arguably best for a single theory of concepts to explain concept acquisition, intentional content, compositionality (combining concepts and thoughts in a productive and systematic way), and categorization (concept application to instances), among others. But it remains unclear whether any single account can do the entire job. Moreover, it should be clear that my focus in this book is on empirical concepts. Machery (2009), however, goes much further than I do, arguing that the theoretical notion of a concept should be eliminated from the theoretical apparatus of contemporary psychology.

Returning to concept possession, it is worth mentioning the oft-cited *generality constraint* on concept possession. The generality constraint is sometimes put as follows: the attribution of thoughts to any organism of the form "a is F" and "b is G" commits us to the idea that the organism should also be able to think that "a is G" or "b is F" (Evans 1982). Moreover, we might suppose that "the content of all propositional attitudes is said to be subject to this constraint" (Toribio 2007, 446). Thus we might also think of the generality constraint as involving a commitment to the notion that the content of beliefs and desires can be recombined with other such states, and perhaps even that the organism can at least make appropriate simple inferences among them.

However, the generality constraint is, by virtually all accounts, a strong necessary condition to place on concept possession. For example, Tye (2006b, 506) calls it a "stronger requirement" than others he considers because it requires that I am capable of thinking *any* thoughts that can be formed by combining a concept with other concepts I possess. This, however, makes all concept users into idealized rational agents capable of combining into thoughts all concepts that one possesses. Furthermore, a great deal of reasoning and inference requires the more sophisticated introspective capacity that, in the HOT model, takes place at the level of *conscious* HOTs.

One might instead opt for the position that possessing a concept C is one of degree, which then plausibly allows for "a partial understanding of C. On this intuitively attractive view, one cannot possess the concept *fortnight*, for example, unless one grasps that a fortnight is a period of time" (Tye 2006b, 506). *There are degrees of understanding a concept*. Of course, we must also be careful not to make concept possession so minimal as to leave us with a trivial notion of conceptualism and thus an uninteresting notion of nonconceptual content.

My own view is fairly minimal, but I do not think it is problematically so. First, I think that possessing a concept C normally involves being able to discriminate instances of C's from non-C's. If a subject S has a concept C, S should be able to differentiate instances of C's from non-C's, at least to some extent. This has more to do with perceptual concepts. Indeed, many psychological tests are based on this basic notion of concept possession. However, mere discrimination does not seem to be quite enough. We can perhaps imagine someone S being able to reliably sort, say, pictures of gibbons from pictures of orangutans even though S does not have a significant concept of gibbons.

S should presumably also be able to *identify* or *recognize* instances of C by virtue of at least one of C's central features. Recognizing gibbons independently would at least show a better understanding of gibbons. Thus, in addition to discriminating instances of C's from non-C's, a subject S should at least be able to recognize or identify instances of C's as having certain features or properties. This also seems to capture what conceptualists have in mind when talking about concepts being *deployed* in experience. Finally, even if we reject the generality constraint, it still seems reasonable to hold that if S has a concept C, then S must be able to at least have some intentional states (thoughts, beliefs, and so on) with C as a constituent. After all, why else would we attribute such concepts to an organism?

But it is again important to recognize that there are degrees of understanding along each dimension, namely, discrimination, recognition or identification, and thought capacity. So I do not agree with those who speak of needing a *mastery* of a concept (Bermúdez 1998). Tye, a nonconceptualist, rightly allows that "the ability to exercise a concept in thought does not require full mastery of the concept" (2006b, 506; cf. Speaks 2005, 378). Let us therefore use the following criteria for concept possession, labeled CONPOSS, though I will raise some further questions later. Together with CON and NC, we will see how to apply CONPOSS to anticonceptualist arguments later in this chapter.

(CONPOSS) Whenever a subject S has an empirical concept C that is applied to some object (or property or relation) in experience *e*, S must at minimum (a) be able (to some extent) to *discriminate* instances of C's from non-C's, (b) be able (to some extent) to *recognize* or *identify* instances of C by virtue of at least some central feature of the objects or properties in *e*, and (c) be able to include the concept C in at least some intentional states that S has.

By "central feature," I do not mean "necessary condition"; I mean a feature that many, if not most, instances of C have. For example, a central feature of tables is that they have four legs. In other cases, it may be a necessary condition, such as the idea that a tiger must be an animal. The same goes for Tye's fortnight example, which is, at minimum, a unit of time. CONPOSS is surely not some trivial notion of concept possession, which renders CON true by definition or NC automatically false.

As we will see in later chapters with respect to animals and infants, this notion of concept possession also has other important implications. As Allen explains: "Philosophers have been tempted by the argument that . . . for example, a dog does not believe there is a squirrel in the tree because it lacks 'the' . . . concept of squirrel. But there is no reason to think that having [that] belief requires that animals have that specific concept, nor that lacking the canonical concept of squirrel means that they lack any concept whatsoever" (Allen 1999, 35–36). We might also borrow the related core idea from the animal cognition literature that the attribution of a concept is justified if evidence supports the presence of a mental representation that is independent of solely perceptual information (Allen and Hauser 1991). In other words, the organism must have a kind of flexibility such that it does not always respond in a fixed way to stimuli.

6.1.3 Subpersonal Nonconceptual Content

Nonconceptualists are often concerned with subpersonal representational states or contents, say, in early visual perception (Bermúdez 1995). This way of applying the notion of nonconceptual content is certainly interesting and important, but it is not my primary concern here. I am mainly interested whether or not there is nonconceptual content *in conscious experience*, that is, on the personal level. This is, after all, primarily a book on consciousness. Nonetheless let us look briefly at a few examples. As we will see, they pose no threat to CON and also reinforce some of the points made in chapter 2.

(1) One example of what appears to be subpersonal nonconceptual content can be found in Marr's (1982) well-known computational theory of vision, according to which there are three levels of visual processing. Marr characterizes the contents of these levels or states in terms of concepts that are clearly not possessed by the average person. For example, there is first a "primal sketch" representation of the visual scene akin to a pixel array. The visual system then generates a "two-dimensional" (2-D) sketch that represents the boundaries of objects and, for example, separates figure from ground. Finally, a 3-D model captures the entire three-dimensional structure of objects. The 3-D image abstracts away from surface textures and information about specific size relationships, which is ideal for object recognition. Marr also uses other notions, such as "zero-crossings," that are similarly not possessed by the average subject.

(2) Raftopoulos and Müller (2006) extensively review the relevant empirical work and argue that what makes the content of a representation nonconceptual is precisely that its content is insulated from first-person access. Such states are thus subpersonal, and this is precisely why their content is nonconceptual. According to Raftopoulos and Müller (2006), conceptual and nonconceptual content are different kinds of content (cf. Raftopoulos 2009a).

(3) Pylyshyn (2007) also uses nonconceptual content in early vision to present an account of perceptual reference. He discusses a number of "pointers" or "indexes" directed at external objects that function like demonstratives. In describing them, he uses the acronym FINST, which stands for "fingers of instantiation." FINSTs "give us nonconceptual access to what I have called a *thing* or *sensory individual*. . . . Because the representation is not conceptual, these sensory individuals are not represented *as* objects or as Xs for *any* possible category X" (56). The idea is that there are "demonstrative pointers" in the early visual system that allow an organism to parse the visual world and so segregate things in space and time. Pylyshyn argues

that this form of reference gains support from our ability to successfully engage in what he calls "multiple object tracking."

(4) It is also worth noting that Fodor (2007) has reached a similar conclusion, namely, that there may indeed be unconceptualized representations, but only in the subpersonal and informationally encapsulated early stages of perception before conscious awareness. He calls such nonconceptual representations "iconic" representations, that is, a representation R where every part of R represents a part of what R represents. Such representations, Fodor insists, are more "picture-like" than sentencelike.

I should say here that a conceptualist might even challenge any of the foregoing accounts on their own merits. For example, a conceptualist might still wonder why OBJECT or SPACE or SURFACE is not applied in Pylyshyn's or Marr's theories of subpersonal content. I will address this more directly in the next chapter. My point here is only that some accounts of nonconceptual content are put forward, in part or in total, as instances of subpersonal representational content. As such, they do not and cannot falsify CON, since they are not aimed at showing that there is nonconceptual content in conscious or personal level experience. Indeed, CON is compatible with what these authors have in mind.

Perhaps more importantly, this discussion again raises a key issue addressed in chapter 2. For example, what is the difference between a mere "informational" or "computational" state and a genuine *mental* state? Recall that I rejected Searle's Connection Principle (CP), according to which each genuine unconscious mental state is potentially conscious. Nonetheless we should not say that merely because a mental state is unconscious that it is permanently subpersonal or automatically lacks genuine representational content. Most unconscious mental states, such as my current unconscious beliefs and desires, are conceptual and potentially conscious. But there still seems to be an important difference between unconscious purely computational or informational states, say, at the level of early visual processing, and the usual assortment of unconscious intentional states. Unconscious informational states might still be representational states in some important way but yet cannot become conscious. The contents of these states can indeed be nonconceptual but also seem entirely restricted to the subpersonal level. Intentional states are unconscious states with conceptual content, such as my current beliefs about the capitals of various countries. As we have seen, one way to mark the difference in question is to hold that purely informational states do not have any level of what Stich called "inferential promiscuity," that is, inferential integration with the rest of the cognitive

system. They are precisely those informationally encapsulated representations, according to Fodor.

In chapter 2, I alluded to the two-visual-systems hypothesis (Milner and Goodale 1995; Goodale 2007), according to which there is the ventral (conscious) pathway in the brain and the dorsal (unconscious motor) pathway.[7] This presented a problem for CP provided that we were willing to treat dorsal pathway representations as genuinely intentional. Moreover, with respect to the Ebbinghaus illusion (the Titchener circles) (fig. 6.1), Andy Clark (2001, 502–505) has emphasized findings that although subjects will consciously perceive, via the ventral pathway, that the circles (represented as discs) are different sizes, there is evidence that, even when subjects were unaware of the illusion, their motor control (dorsal) systems led them to produce a precision grip with a finger-thumb aperture perfectly suited to the *actual* size of the disc (Aglioti, DeSouza, and Goodale 1995; Goodale 2007, p, 622).[8] The main point here is that there is little reason to suppose that any conceptual representations in the dorsal stream are being deployed in the subject's *conscious* experience.[9]

6.2 HOT Theory and Conceptualism

Although I think that CON is independently defensible, it is rarely, if ever, viewed in light of a well-developed theory of consciousness. It seems to me that HOT theory and conceptualism fit together hand in glove. In this section, I elaborate on some of these connections.

6.2.1 Some Basic Connections

Having, in my view, established the HOT Thesis, we can now formulate a more explicit and direct argument for conceptualism as follows:

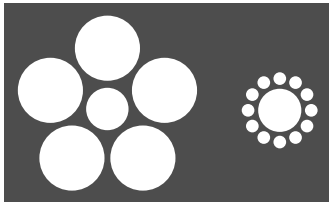


Figure 6.1
The Ebbinghaus illusion/Titchener circles. Despite appearances, the inner circles are the same size.

The HOT-CON Argument

(1) Whenever a subject S has a conscious perceptual experience *e*, one has a HOT directed at *e*.

(2) Whenever a subject S has a HOT directed at *e*, the content *c* of S's HOT determines[10] the way that S experiences *e* (provided that there is a match with the lower-order state).

(3) Whenever there is a content *c* of S's HOT determining the way that S experiences *e*, the content *c* (of *e*) is fully specifiable in terms of concepts possessed by S.

Therefore,

(4) Whenever a subject S has a conscious perceptual experience *e*, the content *c* (of *e*) is fully specifiable in terms of concepts possessed by S.

Notice that the conclusion is identical with CON. CON naturally falls out of the HOT theory. It is also important to keep in mind the key point, embodied in premise (2), that a HOT (and thus its constituent concepts) not only makes the experience conscious but also contains already possessed concepts. As we saw in chapter 4, there would still need to be a match between the LO and HO conceptual contents. However, for reasons that will become clear later in the chapter, it will be necessary to refine this view. Premise (3) simply states the obvious fact that, according to the HOT theory, a HOT is constituted by possessed concepts. Whatever is the case with perception, it is relatively uncontroversial that thoughts are constituted by concepts. Several other points are worth emphasizing here.

First, recall the following quote from Rosenthal where he says that it is "plausible that peeling away that weakest HOT would result, finally, in its no longer being like anything at all to have that sensation" (2002a, 414). Not only do our concepts color the very experiences we have, but removing all of them would eliminate the experience itself. This certainly sounds like an indirect endorsement of CON. Rosenthal argues that if we systematically remove *all* the relevant concepts involved in having a conscious auditory experience (such as the "sound of a woodwind"), then there would no longer be the conscious experience at all. If so, then there would seem to be no room left for nonconceptual conscious experience.

Second, we are all familiar with the phenomenon of "seeing-as" whereby one subject, perhaps with more knowledge, might see an object as a tree, whereas another person might only see it as a shrub. The same is true for "hearing-as," for "tasting-as," and more generally for "representation-as." This phenomenon is particularly noticeable in cases of perceiving ambiguous figures, such as the well-known vase–two faces picture. I examine these cases in more detail in the next subsection, but here is another relevant

quote from Rosenthal: "There being something it's like for one to have a sensation of red, for example, will consist in one's being [aware] of that sensation *as* a sensation of red; similarly for other kinds of cases" (2005, 4).

Third, I have noted my sympathies to various Kantian theses that are also relevant here.[11] For example, we encountered the Kantian idea that we first passively receive data via our senses in what Kant (1781/1965) calls our "faculty of sensibility." Some of this information rises to the level of unconscious mental states but will not become conscious until the faculty of understanding operates on them via the application of concepts. We can understand concept application in terms of HOTs directed at the incoming information. Thus I consciously experience the green table *as a green table* partly because I apply GREEN and TABLE (in my HOTs) to a lower-order unconscious state via my visual perceptual apparatus. More specifically, I have a HOT such as "I am seeing a green table now." It takes the cooperation of both the sensibility and understanding to produce conscious experience.

Fourth, recall from chapter 2 the discussion about Fregean content and modes of presentation. The contents of conscious states include both the Russellian and Fregean variety. Representationalists typically have in mind Russellian contents, but they are not normally thinking in terms of the HOT theory of consciousness. An advantage of the HOT theory is that it can naturally explain how conscious states embody both kinds of content while retaining its reductionist credentials. So the content of, say, an *unconscious* first-order visual perception is typically Russellian, but the content of the complex conscious counterpart is also Fregean. If there is a match between M and MET, the MET will also tell us the way that the objects of first-order states *are presented to the subject*. We might say that the *mode of presentation* is determined by the HOT's content. Thus a "whole" conscious state can be analyzed in a way that accommodates both kinds of content. Nonetheless such a view is still reductive because what accounts for the Fregean content in a first-order conscious state is still itself unconscious. This move is not available to FO representationalists because there is only one level of mental content.

Fifth, keeping in mind my earlier discussion of subpersonal nonconceptual content, it should be noted that according to HOT theory, it is only when those unconceptualized subpersonal representations are taken farther upstream in the cognitive system that a representation gains a level of conceptualization. And when these conceptualized representations become conscious, there are HOTs directed at them. It is precisely when a first-order state becomes targeted by a HOT that the state becomes conscious.

Sixth, one way the conceptualism–nonconceptualism debate is characterized is in terms of the relationship between perception and thought. It seems to me that the HOT theory, and especially the WIV, can be informative in this regard. According to the WIV, an important intimate relationship exists between perception and thought. Indeed, a thought is always built right into any first-order conscious perceptual state. Thus not only does HOT theory fit well with conceptualism, but conceptualism can better be understood in terms of a version of HOT theory. HOT theory *explains why* conscious perceptions are, and must be, conceptualized. There is, we might say, an implicit *judgment* about a representation R when R is conscious.

Seventh, a conceptualist should also recognize that when we have a first-order perceptual state, we do not normally *consciously* apply concepts to our experiences. Rather, they are presupposed in the experience. We might say that our concepts are normally unconsciously applied to the incoming data about the world. This fits well with HOT theory because HOTs are themselves unconscious in such first-order cases. When I have a conscious experience of a tree and thus deploy TREE, I am not consciously thinking about applying TREE at the time.

Finally, recall the view that HOT theory should arguably be understood as a necessary truth in the context of responding to the hard problem of consciousness (sec. 4.4). I think the same sort of claim should be made about CON. Thus one way to put it is that the conceptualist is committed to the following modal claim:

(MODAL-CON) Necessarily, for any two subjects, *S*1 and *S*2, and objects (or properties or relations) *o*1 and *o*2, *o*1 is represented in experience *e*1 differently than *o*2 is represented in experience *e*2 only if *S*1 and *S*2 possess and apply distinct concepts, *C*1 and *C*2, for *o*1 and *o*2 respectively in experiences *e*1 and *e*2.

Thus, speaking more loosely, the conceptualist holds that it is impossible for there to be representational or perceptual differences without conceptual differences. That is, it is never the case that S1 has and applies the same concepts as S2 but S1 and S2 have different conscious perceptual content. Alternatively, it is never the case that S1 has the same conscious perceptual content as S2 but S1 and S2 have or apply different concepts. And this is precisely what opponents of conceptualism deny. I am not claiming that conceptualism entails HOT theory. However, adopting HOT theory can be a good strategy to pursue if one wishes to shed light on conceptualism within the framework of a theory of consciousness. On the other hand, it seems

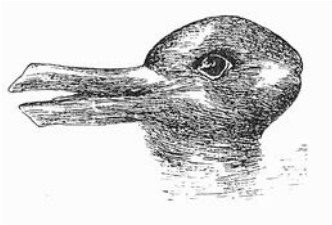
to me that any HOT theorist is committed to conceptualism or least has an importantly close relationship to it.

Upon closer examination, however, it is perhaps not entirely clear that Rosenthal himself holds CON. He has not written explicitly on the subject and never, for example, describes "qualitative character" as a kind of nonconceptual content. Still, let us recall his homomorphism theory, which says that qualitative character is a matter of the ways in which sensory states resemble and differ from one another in a way that mirrors the resemblances and differences between perceptible properties. Some interesting subtleties do arise. For example, as we saw in chapter 4, Rosenthal ascribes qualitative properties to unconscious sensory and perceptual states. He sees these properties as *representational, but not intentional* in the way that thoughts and beliefs are. So he treats intentional content and qualitative character as different properties.

Thus Rosenthal might deny CON if it is taken to mean that sensory and perceptual states, *conscious or not*, have *only* conceptual (or intentional) contents, because he holds that sensory and perceptual states have *other* representational features, namely, their qualitative character. According to Rosenthal, there are unconscious first-order *sensory* or *qualitative* states. At the least, then, Rosenthal might be taken as propounding yet another form of *subpersonal* nonconceptual content in terms of what he calls "qualitative character," especially since he reserves the term "content" for *intentional* content. He has also distinguished between "sensing" and "perceiving" such that sensing "has no conceptual content" (2004, 20), whether first order or higher order. But could we interpret Rosenthal as holding that nonconceptual content is present in *conscious* experience, that is, on the personal level? Perhaps one could interpret his theory in such a way if one takes his notion of "content" broadly so as to include what Rosenthal calls "qualitative properties." It is, after all, precisely these mental properties that we are aware of when we have a conscious perceptual state. In any case, if Rosenthal is not a conceptualist, then I think he and other HOT theorists *should* endorse CON given the HOT-CON argument and the other connections described thus far. Given Rosenthal's frequent reference to Sellars in developing his homomorphism theory, perhaps Sellars is also not quite the ally that McDowell thinks he is.

6.2.2 Ambiguous Figures

Related to the notion of "seeing-as" is the phenomenon of Gestalt switching, whereby one shifts from seeing an object or image in one way to seeing it in another way. Readers will likely be familiar with ambiguous figures



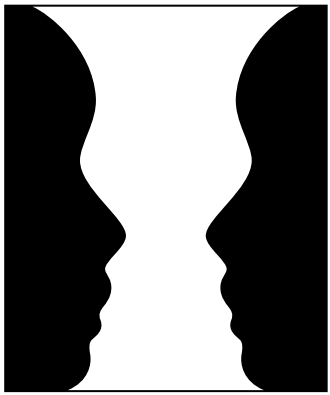


Figure 6.2
Two ambiguous figures: the duck-rabbit and the vase–two faces.

such as the duck-rabbit and the vase–two faces (figs. 6.2a–b). In these cases, it is impossible to see both figures simultaneously. For reasons that will become clear, I think that both conceptualism and HOT theory have an advantage over other representational theories when it comes to accounting for experiences of this kind.

MacPherson (2006) argues in great detail, and fairly convincingly, that at least some of these experiences serve as counterexamples to *non*conceptualist FOR theory, such as the theories of Tye and Dretske (though MacPherson is otherwise more inclined to support NC than CON). The gist of MacPherson's argument is that it does not seem possible to account for all phenomenal differences in Gestalt switching purely in terms of differences in nonconceptual content. It is unclear what nonconceptual differences between the two experiences could explain the change in one's phenomenological first-person experience. Thus we have multiple counterexamples to standard nonconceptualist FOR. Ambiguous figures seem to undermine the central idea that the nonconceptual content of visual

experiences is identical with the phenomenal character of those experiences. These kinds of visual experiences seem to represent properties in addition to the unchanging properties such as color, shape, positions, and sizes. MacPherson does, however, leave open the possibility that changes in nonconceptual content *could* account for *some* cases of Gestalt switching, but not all. It would be better, however, to have a uniform account of all such experiences.

But MacPherson never really considers a natural HOTish solution to the problem. For example, we might first say that when one sees the vase–two faces *as a vase*, one is applying VASE in one's visual experience. When one switches and sees the figure *as two faces*, it would seem that FACES is applied in that experience. Tye and many others seem to concede this much. One is clearly categorizing and experiencing different objects in each case, which involves concept application. At the least, the faces are much more phenomenologically prominent when that concept is applied, and the vase is more prominent in one's experience when that concept is deployed.

So a supporter of CON and HOT theory has the ready reply, for example, that when one sees the vase–two faces *as a vase*, one is applying VASE in a HOT to one's visual experience. *Mutatis mutandis* for other ambiguous figures. The change in conceptual content explains the difference in conscious experience. We might say that there is a *judgment* containing, say, the concept VASE embedded within one's experience of the picture when it is experienced as a vase. This judgment is a HOT to the effect that "I am now seeing a picture of a vase." It is crucial to recognize that the applied concept in question certainly seems to be part of the very experience itself. Tye had previously attempted to divorce the phenomenal experience from the concepts involved: "What happens in cases like these is that one has a sensory representation whose phenomenal content is then brought under the given concepts. *Still, the concepts do not enter into the content of the sensory representations and they are not themselves phenomenally relevant*" (1995, 140; italics mine). I am frankly not sure what to make of this last sentence, especially the claim that the concepts are not themselves phenomenally relevant.[12] Presumably he means something like "VASE is applied to, and thus *causes* a change in, the experience, but VASE is not a *constituent* of the experience itself." I disagree; it seems to me that when I perceive it as a vase, VASE is *part of* the conscious state's content itself. Surely a subject is representing the ambiguous figure as a vase when perceiving it that way. That's what seems to generate the problem in the first place. Moreover, if we hold something like the WIV, then we have a natural way of understanding just how VASE is incorporated into the very conscious state itself, that is, as a concept in the

MET. As we have seen, the meta-psychological judgment in question is part of the overall conscious state itself.

Despite her critique of nonconceptualist FOR, MacPherson doesn't think that one must appeal to the conceptual content of experience to account for Gestalt switching. She offers three main reasons. I will reply to each:

(1) "Not all changes in judgment appear to lead to the special changes that occur in perceptions of ambiguous figures" (MacPherson 2006, 91). For example, there is no Gestalt switch in seeing a titled "A" in the way that there is in, say, the square/tilted diamond case. However, it seems to me that the reason for this is that there is no distinct TILTED A concept in the way there clearly are two distinct concepts applied in the square/titled diamond case. As MacPherson recognizes, the titled A case is more like a mental rotation task than like seeing a single ambiguous figure in two different ways. Indeed, the titled A is not an ambiguous figure at all.

She then notes examples of optical illusions that persist in spite of our clear conceptual knowledge or judgments, such as the Müller-Lyer illusion (fig. 6.3).

The lines in the Müller-Lyer illusion continue to look unequal in length even when one knows that they are equal. Thus possession of these concepts does not alter the conscious perception in this case. But it is unclear why her opponent must accept the extremely strong claim that *any* or *all* concepts possessed by a subject S must bring about perceptual changes. As we saw earlier, there may indeed be some informationally encapsulated nonconceptual subpersonal content that causally contributes to some conscious perceptions. Indeed, these representations are likely too early in the visual process to be mental representations at all. Some standard illusions are indeed good examples of this phenomenon, but it does not follow from this that the *conscious experience itself* has that nonconceptual content. The main point is that when one does consciously perceive an

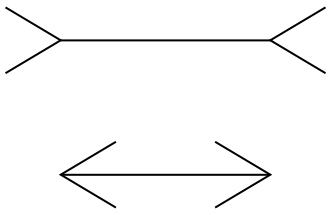


Figure 6.3
The Müller-Lyer illusion. Despite appearances, the horizontal lines are the same length.

object O, one sees it as one's concepts are deployed in *that* conscious perception of O.

(2) MacPherson then argues that Gestalt switching can often happen without one's control, which "seems to suggest that the visual system has a certain autonomy . . . that is quite unlike ordinary judgment" (2006, 92). I agree that this happens, but it is not inconsistent with the view presented here. One need not consciously or voluntarily apply concepts to a visual image to experience a Gestalt switch. All that matters is the application of the concept. Indeed, in the WIV (as in HOT theory), the MET is unconsciously applied in cases of first-order perceptions. But MacPherson is right that Gestalt switches can sometimes be voluntarily controlled with a "certain autonomy." Perhaps it is partly a matter of attentively focusing on some areas of the ambiguous figures. I would only suggest here that when voluntary control is present, there may temporarily be a shift to conscious HOTs whereby we are consciously thinking about the relevant concepts involved while trying to see the ambiguous figure one way or another. This process helps to bring about a different first-order experience.

(3) Finally, one might even resist the claim that perceiving differences in ambiguous-figure experiences must involve applying different concepts, such as in cases of infants or children (95–96). MacPherson acknowledges that she knows of no study showing that creatures lacking the relevant concepts can still perceive Gestalt switches, as in the duck-rabbit example. Yet she doesn't want to rule out this possibility. Of course, part of the point of this entire chapter (and the next) is to defend CON and show that it is consistent with the Infants Thesis. But in discussing this possibility, MacPherson admittedly uses a more sophisticated "high-grade" notion of concept possession than we have in CONPOSS. Moreover, even if an infant or other creature can perceive a Gestalt switch without DUCK and RABBIT, it must surely still be the case that she is differentiating and identifying the figures by virtue of *some* concepts, such as ANIMAL FACING LEFT WITH A BIG MOUTH, as opposed to ANIMAL FACING RIGHT WITH LONG EARS. Perhaps it is even possible to have a Gestalt switch, similar to the vase–two faces case, but with two nonsense figures on either side of another figure for which we do not have distinct concepts. This is perhaps a legitimate possibility, but it seems to me that we would still be able to point to some concepts that apply only to each figure.

In any case, we can thus also respond to MacPherson's more general doubt that there is any "naturalistic theory of representation . . . that can predict that there will be a representational difference between experiences of ambiguous figures" (2006, 109). I suggest that HOT theory, coupled with a closely related conceptualism, does indeed have the resources to predict

and explain the representational differences in question. Once again, the advantage over FOR theory here is partly due to the additional layer of representation within a first-order complex conscious state. This allows a HOT theorist to acknowledge that there is indeed some unchanging but ambiguous sensory input in these cases while still using the concepts in the HOTs to account for the experiential differences in question, that is, *the way* that such information is consciously perceived.[13]

Some empirical support exists for the foregoing line of argument. Research on hybrid and bistable ambiguous figures provides quantifiable evidence for the ways that categorization and top-down attention influence perception (Schyns and Oliva 1999; Bonnar, Gosselin, and Schyns 2002; Özgen et al. 2005). Bonnar, Gosselin, and Schyns (2002) use Salvador Dalí's painting *Slave Market with the Disappearing Bust of Voltaire*, which contains an ambiguous image that can be seen either as the heads of two nuns or the eyes of the bust of Voltaire. Schyns and Oliva (1999) focus on the perception of faces. Once again, categorization is closely bound with perception. The selection of certain categorical information, such as task-dependent factors, sound cues, and expectations, can modify the perception of the input.

6.2.3 Associative Agnosia

We can also view the relationship between conceptualism and HOT theory through an objection that might be raised against both views.

Some have suggested that visual agnosia presents a problem for CON (Bermúdez 1998, 79–82; A. Smith 2002, 112–113). Their reasoning suggests a similar objection to HOT theory. Visual agnosia, or more specifically *associative agnosia*, seems to be a case where a subject has a conscious experience of an object without any conceptualization of the incoming visual information. There appears to be a first-order perception of an object without the accompanying concept of that object (either first- or second-order, for that matter). Thus its "meaning" is gone and the object is not recognized. In short, it seems that there clearly can be conscious perceptions of objects without the application of concepts, that is, without recognition or identification of those objects.

Let's first distinguish between *apperceptive* visual agnosia, cases where "recognition of an object fails because of an impairment in visual perception," and *associative* visual agnosia, cases "in which perception seems adequate to allow recognition, and yet recognition cannot take place" (Farah 2004, 4). I am concerned with associative agnosia, instances described as having a "normal percept stripped of its meaning" (Teuber 1968). So, for example, a patient will be unable to name or recognize a whistle.

Associative agnosics are not blind and do not have damage to the relevant areas of the visual cortex, as is the case with blindsight patients. In addition, associative agnosics tend to have difficulty in naming tasks and with grouping objects together. Unlike in apperceptive agnosia, there seems to be intact basic visual perception; for example, patients can copy objects or drawings that they cannot recognize, albeit very slowly. But the deficit in associative agnosics is more cognitive than in patients with apperceptive agnosia. Patients will also often see the details or parts of an object, but not the "whole" of the object at a glance. The main point is that the phenomenal character of associative agnosics has changed in a way that corresponds to a lack of conceptual deployment. It is, however, important to recognize that associative agnosics still do have the relevant correct concept because they can apply it to *other* modalities. For example, a patient might identify a whistle by sound but not by sight.

Nonetheless, upon closer examination, I believe that associative agnosia is perfectly compatible with both conceptualism and HOT theory. Let's begin with conceptualism.

First, there is nothing in conceptualism itself that says a subject S must always apply the *correct* object concept to outer objects or properties. This would be a level of mind-to-world infallibility that few, if any, would endorse. Moreover, CON simply says that the content of the *experiences* in question needs to reflect concepts possessed by the subject. Thus, in cases of associative agnosia, the problem is simply that, in cases of vision, the mechanism by which the appropriate concept is triggered is defective for some reason. So while one might be said to have a conscious visual perception *of a* whistle in some sense, the patient is not experiencing the object *as a whistle*.

Second, Farah (2004, chap. 6) makes it clear that associative agnosics do *not* really have normal perception from the subject's phenomenological point of view, but merely with the elimination of concept deployment. They do not perceive things *just like others* save for a recognitional element. For example, although they can produce excellent drawings of unrecognized objects, the process is abnormally slow and narrowly focused on one part of the object at a time. Farah tells us that "close scrutiny invariably reveals significant perceptual abnormalities in agnosic patients" (2004, 74). Moreover, associative agnosics also have some difficulty in perceptual tasks involving odd and otherwise meaningless shapes. Thus the deficit seems to be more of a global perceptual problem than it might initially seem.

Third, as long as the visual perception's content falls under concepts possessed by S, then it is compatible with conceptualism. Thus, in the case

of a whistle, it seems better to describe the content of S's visual experience as directed at, say, a silver, roundish object. So a conceptualist can still hold that some *other* concepts are present in such abnormal cases, which precisely reflects the way that S is experiencing the object. S is then still deploying concepts like SILVER and OBJECT, and those concepts *are* represented in S's experience. Thus Smith is mistaken in holding that agnosia is a case where one "can perceive something and *wholly* fail to classify it, fail to perceive that it is any particular *kind* of thing at all" (A. Smith 2002, 112; first italics mine). Smith also misses the point of conceptualism when he says that it is "absurd" and "incoherent" to suppose that "you have to recognize . . . everything you perceive" (112). Conceptualism is not committed to this absurd view. Rather, it holds that the *content of your experience* must be matched by concepts you possess. Smith might reply that there is still surely *some sense* in which agnosics (and the rest of us, for that matter) *see* things that we do not recognize. Perhaps this is so in some nonintentional sense of "see" whereby one's eyes are pointed at or looking at some unrecognized object. But the issue is what one is consciously experiencing the object *as* at that time. The content of that visual perception is still entirely structured by concepts possessed by the subject. Turning now to HOT theory, a similar reply is possible.

If HOT theory is true, there must be *some* HOT for there to be a conscious state at all. So a HOT theorist can hold that a nontypical and less-robust HOT is present in such abnormal cases, which reflects the way that the patient is experiencing an object, such as a whistle or paintbrush. If one experiences object O only as having certain parts or fragments, then those concepts will be in the relevant HOT. But there is no reason to suppose that HOTs are *entirely* absent in these cases. As we saw in the whistle case, perhaps only SILVER, ROUNDISH, and OBJECT are applied to O. As long as that is how the agnosic *experiences* the object, then HOT theory is left unthreatened.

What most complicates the matter for HOT theory has to do again with the difficult problem of misrepresentation, which I addressed initially in chapter 4. I primarily used Levine's (2001) example of the red/green diskette case, but we saw that some of the options available to the HOT theorist were not very attractive. For example, choosing between the LO state and HO state to determine the color experience is problematic. Thus I argued that the best reply to this objection is to construe the LO and HO states as part of a single integrated state, so that misrepresentation cannot occur *and* still result in a conscious state. That is, a conscious state only results when the LO and HO concepts match. Indeed, this was a major motivation to move from standard HOT theory to the WIV.

Associative agnosia also highlights and supports my contention that the "lower" order state is inextricably bound up with the "higher" order state. As we saw earlier, this strategy is supported by Farah's analysis. There really is no raw and completely intact visual perception that can be separated from the "meaning" or "recognition" of the object. This point is reinforced by Riddoch and Humphreys (1987), who discuss what they view as a common and more specific kind of associative agnosia under the heading of *integrative agnosia*. They observe that patients often guess the identity of objects based on a single local feature or parts of objects. Agnosics often mistakenly guess at the identity of an object (such as calling a baby carriage a "bicycle") on the basis of a single feature (such as wheels). In light of the misrepresentation issue, perhaps we should also understand associative agnosia as a problem of "impaired integration of local shape parts into higher-order shapes" (Farah 2004, 78).

In addition, it might be helpful to think of what happens in associative agnosia as analogous to perceiving an ambiguous figure. That is, the incoming visual information about an object for these patients is ambiguous and so is accompanied by an atypical or weaker HOT that does not apply WHISTLE. Following up on this example, WHISTLE is not in the HOT for the agnosic, but SILVER and ROUNDISH are. Thus the visual experience reflects only those concepts analogous to a kind of faulty gestalt perception. Relative to HOT theory, this suggests it takes the integration of both the LO state and the HOT to produce a coherent and accurate visual experience of an object. Otherwise a loss of "global" or "gestalt" perception occurs. But this is not a problem for HOT theory as such, or at least for the WIV, since we cannot really separate the LO and HO states. What is missing for the associative agnosic, however, is the ability to have a gestalt switch to *recognizing the overall object* that would require applying WHISTLE.

These responses are consistent with my prior discussion of the misrepresentation problem in chapter 4. Nonetheless my solution does require some further refinement. We might hold that associative agnosia is simply an unusual case where the typical HOT does not *fully* match up with the first-order visual input. That is, we might view associative agnosia as a case where the "normal," or most general, object concept in the HOT does not accompany the input received through the visual modality. There is a *partial match* instead. A HOT might *partially recognize* the LO state. So associative agnosia would be a case where the LO state *could* still register a percept of an object O (because the subject still does have the concept), but the HO state is limited to some features of O. Bare visual perception remains intact in the LO state but is confused and ambiguous, and thus the agnosic's

conscious experience of O "loses meaning," resulting in a different phenomenological experience.

It may still seem that this way of handling associative agnosia is at odds with my earlier treatment of the misrepresentation problem. Notice, however, that in the foregoing cases there is no mismatch in the sense of there being *incompatible* properties represented in the LO and HO states. In Levine's red and green diskette example, the case cannot be both entirely red and green. On the other hand, in the ambiguous-figure case, ambiguous LO content can be "recognized" by a HOT in two incompatible ways (as a vase or two faces), resulting in two very different perceptual experiences. But the same LO representation *does* nonetheless represent both a vase and two faces. A similar reply applies to the agnosia case: ambiguous and thus *more general* LO content is accompanied by more specific HO content. So there is an important difference between Levine's case and the scenarios where the LO content is ambiguous or more general than the HO content. HOTs can thus serve the purpose of narrowing down the *conscious* perceptual content in these cases. The HOT is recognizing only part of what is present at the LO level. According to the WIV, however, if there were no LO state at all or incompatible concepts in LO and HO, then there would be no conscious state for the reasons we saw in chapter 4.

Thus we should modify the parenthetical statement made in premise (2) of the HOT-CON argument, namely, that "whenever a subject S has a HOT directed at *e*, the content *c* of S's HOT determines the way that S experiences *e* (provided that there is a match with the lower-order state)." It should now read:

(2′) "Whenever a subject S has a HOT directed at *e*, the content *c* of S's HOT determines the way that S experiences *e* (for whatever *full or partial* conceptual match exists with the lower-order state)."

This does not alter the validity of the HOT-CON argument, since premise (2′) can also be substituted into the beginning of premise (3). We will see later that further refinement will emerge from the discussion of fine-grained experiences.

Two final points: First, it is worth noting that some of the perceptual difficulties facing associative agnosics are similar to *prosopagnosia*, which is "the inability to recognize faces despite intact intellectual functioning and even apparently intact visual recognition of most other stimuli" (Farah 2004, 92). Prosopagnosics "often speak of seeing the parts [of a face] individually and losing the whole or gestalt" (94).

Second, we might also view the overall problem here as a breakdown in the unity of consciousness, that is, in terms of a lack of the binding of *features* of objects into perceptual wholes. Many attempts to understand the visual system or visual processing (in the brain) in terms of binding use levels of representation and integration. The suggestion here is that we can also understand the visual object recognition process in terms of the proper integration of a LO state and an appropriate HOT.

In any case, I think that associative agnosia and ambiguous-figure experiences can be neatly accounted for by HOT theory and conceptualism. Such cases reinforce the explanatory power of HOT theory and the importance of matching HO concepts to LO input. These cases also show the advantage of understanding normal visual perception as genuine *integrations* of LO and HO states, as is the case with the WIV.

6.3 The Richness of Conscious Experience

6.3.1 The Main Argument

One attempt to support NC, and thus falsify CON, starts with the premise that many perceptual experiences are extremely rich in content. It seems, for example, that we can *simultaneously* experience a complex visual scene, such as a landscape, and it seems implausible to suppose in those cases that the subject deploys concepts for every object (and property and relation) that the experience represents. Numerous objects, shapes, and colors are represented, not to mention the relations between them. Unlike beliefs and thoughts, which solely have conceptual content, perceptual experiences represent in a way that goes well beyond one's conceptual capacities. In short, we can have perceptual experiences that outrun our conceptual capacities. As Dretske might put it, a perceptual experience can carry much more information than one is able to conceptualize.[14]

6.3.2 The Deflationary Strategy

A conceptualist can respond to the richness argument in a number of ways. She might initially challenge the central premise that conscious experience is very rich. The claim is that, contrary to initial intuitions, we really do *not* consciously experience very much at any given time. The rationale for this strategy comes from several different and, in my view, compelling sources.

First, recall the distinction between focal and peripheral awareness (and thus perception). As we have seen in previous chapters, there seem to be many cases of conscious perception where we only have focal awareness of a small portion of our visual field or a limited attentional focus in an

auditory experience. In the visual case, this is supported by experiments showing that it is only the center of the retina that has a high density of cones with high acuity. This contrasts with the periphery (or parafovea) of the retina, which allows for much lower resolution.

Consider also Dennett's case of the Marilyn Monroe wallpaper, where you walk into a room with wallpaper containing hundreds of her portraits. Your initial sense might convince you that you are seeing hundreds of identical Marilyns. But are you really? Dennett persuasively argues that the real detail is not in your head but in the world. We simply assume that all the pictures are of Marilyn Monroe; that is, our brains "fill in" the rest of the scene. We thus mistakenly assume that all of the Marilyns are represented in our experience (Dennett 1991, 354–355). This likely occurs often when we experience a number of similar-looking objects at the same time, unless one object is so different as to "pop out" in the experience. You obviously do not focus in (or foveate) on each and every portrait. Indeed, it would seem that you are only peripherally aware of the vast majority of portraits at any given time. It is unlikely that you would notice if, say, six or seven of the portraits were altered to contain portraits of another blonde female.[15]

Second, even *within* focal awareness, there is support from two striking empirical results in visual perception. Some argue that the phenomena of *inattentional blindness* and *change blindness* call the richness of experience into doubt (Noë 2004, 2007; Blackmore 2004, chap. 6).

Inattentional blindness occurs when normal subjects do not notice objects in their visual field while their attention is occupied by a specific task (Mack and Rock 1998). Perhaps best known is the video of a group of people passing a basketball among themselves, and observers are asked to count the number of passes within the group. Many observers do not even notice that someone dressed in a gorilla suit walks right into the center of the scene, pounds his chest, and then walks away (Simons and Chabris 1999). This is a shocking result to many of those tested. It actually seems to me that these cases differ from the usual peripheral awareness, such as was discussed in the previous chapter or similar to the Marilyn wallpaper. Those who do not notice the gorilla tend to report that they are not conscious at all of the gorilla, as opposed to being peripherally conscious of it.

Similar cases can be found when a magician performs sleight of hand even within an extremely narrow focal area of one's visual field, such as when a magician performs a card trick. And, perhaps most surprisingly, this is accomplished even when we know that the magician is trying to fool us! By means of subtle diversions of attention and other techniques, magicians

are often able to cause inattentional blindness in their audiences (Martinez-Conde and Macknik 2008).

Finally, there is evidence that something like inattentional blindness generalizes to members of an attended category in ordinary perceptual contexts (Koivisto and Revonsuo 2007). An unexpected stimulus conceptually related to the observer's current interests, such as words or pictures of animals or furniture, is likely to be seen (even if the representational format is different), whereas unrelated unexpected stimuli are unseen. For example, when subjects attended to pictures, they detected the unexpected word stimulus more often when its semantic category was congruent with that of the attended pictures (94 percent) than when its category was incongruent (41 percent). The important point is that the "meaning" of the stimulus must have been activated *preattentively*, that is, before the stimulus was detected. Meaning can indeed shape seeing.

Change blindness, on the other hand, occurs when normal subjects fail to notice what would seem to be an obvious change in some object or scene (D. Simons 2000). Even in cases where one can compare pictures side by side, subjects often take an extremely long time to notice the change. This suggests that we really do not have a very detailed sense of everything in our visual field. Examples here might include a change in one of the items on a desk or a difference in the number of windows of a building. People often greatly overestimate their ability to detect such changes (Levin 2002).

Moreover, it is well known that there are many quick and jerky eye saccades (or movements) when a subject is looking at a scene or picture. Our eyes dart around in ways that subjects are unaware of, and in the case of change blindness, there is a clear searching of the pictures to find the difference in question (see Tye 2006b, 512, for one example). So it is doubtful that all or most of a visual scene is really *simultaneously* perceived in a way required for the richness argument. If the alleged richness comes from shifting one's attention to numerous different places in a picture or scene over time, then this is hardly a major problem for the conceptualist.[16]

Thus I am sympathetic to the notion that our perceptual experiences are not (or, at least, *very often* not) as rich as they might seem. But even if we have *some* very rich perceptions, that still does not, by itself, seem to refute CON. One might also question the overall validity of the argument even if one is willing to concede the premise that at least some perceptual experiences are rich (Chuard 2007). For example, with Dretske in mind, Chuard rightly explains that even if "the digital content[s] of judgments and thoughts are conceptual, it doesn't follow that the analog content of experience isn't conceptual" (2007, 33). And nothing in the richness

argument really supports the contention that the digital–analog distinction must map onto the conceptual–nonconceptual distinction.

6.3.3 Memory-Based Arguments

Some oppose conceptualism based on arguments exploiting the connection between perception and memory. Each argument is designed to show that one's initial experience must have been richer than the concepts possessed at the time. There are a number of arguments along these lines, but I focus on the following three cases:

Case 1: Sperling's experiment. Perhaps the most prominent argument is based on Sperling's (1960) well-known work on so-called iconic memory. Some authors argue that Sperling's experiments should lead us to reject conceptualism (Tye 2006b; Dretske 1981). More specifically, the idea is to show that we are normally unable to conceptualize everything that we see.

The experiment begins by showing subjects an array of letters in the center of one's visual field for fifty milliseconds, such as an array composed of three rows of four letters each (fig. 6.4).

A visual image of the stimulus was found to persist for 150 milliseconds after removing the stimulus. Subjects were then asked to report what they saw under two different conditions. In condition one, subjects were asked to identify as many letters as possible. In condition two, subjects were asked to identify letters in a single row, albeit after the offset of the stimulus. Sperling found that in condition one, subjects could identify at most only one-third of the twelve letters, and in condition two they could still typically report correctly on at least three out of four. Some conclude from this that one's sensory memory (which does fade quickly) still preserves information

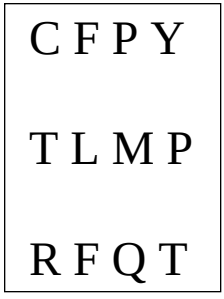


Figure 6.4
An example of an array of letters used in Sperling's experiments.

about the letter shapes in *all* rows although subjects cannot report on all the information. In condition one, it may be that the act of reporting just takes too long and the sensory memories have faded. In condition two, the sensory memory is still available enough to be able to report on most or all letters in a row. Thus the idea seems to be that "at each moment, the visual experiences humans undergo are at least as rich representationally as the sensory memories. . . . [And so] they represent more than their subjects are *capable* of judging to be present" (Tye 2006b, 513). The idea is that subjects are able to perceive more letters than they are able to conceptualize (e.g., identify or recognize).

Case 2: The dice case. Martin (1992) presents the case of Mary, who plays an unusual game of dice. One die is eight sided (octahedral); the other is twelve sided (dodecahedral). But Mary does not have the concepts DODECAHEDRON or OCTAHEDRON. As a matter of fact, Mary is somewhat deficient in basic geometry and doesn't even like counting past five. Although Mary does discern a perceptual difference between the dice due to colored spots, she treats them both as "many-faced shapes." Martin explains that at a later time, Mary does acquire the concepts in question. She then recalls her experience playing the game and realizes that one of the dice was dodecahedral. Martin argues that the content of Mary's original experience of the twelve-sided dice must have been nonconceptual (at least in part) because it is only later that she acquires DODECAHEDRON, and her current memory experience is of her throwing a twelve-sided die. According to Martin, the die still initially appeared to her as a twelve-sided figure.

Case 3: The mustache case. Following up on an example from Dretske (1993) that has also been used against HOT theory, Byrne (1997, 113–114) discusses a case where I see Fred on Monday and then on Friday. Suppose that Fred had a mustache on Monday but shaved it off by Friday. Suppose that I do not notice that Fred has shaved, but surely I *saw* the mustache on Monday and did not see it on Friday. I am thus not aware that these experiences differ. But then there must be differences in my visual experiences of the world of which I am not conscious and thus do not conceptualize. Byrne rightly points out that this kind of case does not refute HOT theory because, as Dretske himself acknowledges, even if there were a conscious difference *between* the Monday and Friday visual experiences, "it does not follow from this conscious difference in my experiences that I am conscious *of* the difference [itself]" (114). That is, I can have different HOTs, and thus experiences, on each day without being aware *that* my Friday HOT differed from my Monday HOT. As Dretske would put it, there is *object*-awareness in each separate case without *fact*-awareness of the difference in question.

Nonetheless Chuard (2007, 38–39) presents a similar case on behalf of the nonconceptualist titled "The General's Moustache," whereby you are recalling an earlier meeting with a general. You later recall that the general previously had a mustache by thinking about what he had looked like, even if you did not notice the mustache at the time. This suggests that what was remembered was still (nonconceptually) represented in the initial encounter with the general. As in the case of Fred, a nonconceptualist can then urge that a subject might not conceptualize some of the things represented in her experience.

Despite the prima facie plausibility of these three cases, I remain puzzled as to how exactly they are supposed to refute CON and thus support NC. The main problem is that there are many doubtful background assumptions with regard to the connections between perceiving, noticing, reporting, and conceptualizing. There are also highly questionable assumptions about memory itself.

Reply 1: With respect to the Sperling experiment, there seems to be an assumption that "we must remember, or at least store in short-term memory, everything that we have previously conceptualized." Why accept this? It may be that the subject *did* conceptualize each of the letters in the array *at the time* of initial exposure, and then simply failed to remember them all later. As Chuard puts it, the "argument assumes that what is conceptually identified exactly corresponds with what is stored in short-term memory" (Chuard 2007, 37). Tye tells us that subjects *believe* that there are twelve letters in the array and indeed see all twelve (2006b, 513). But it might also simply be that there isn't enough time to conceptualize (i.e., explicitly identify) each letter though the subject does at least remember *that* there were twelve letters. Neither of these possibilities is inconsistent with CON. I am inclined to think that not all twelve letters were conceptualized or experienced initially.

Moreover, the argument assumes that whatever letters are initially conceptually identified must correspond exactly to what the subject is able to *report* later. This is an even more problematic assumption and adds another step beyond the initial visual experience. Those who use Sperling's experiment as a basis for NC seem to assume that there is a seamless and infallible transition from a visual perception to a memory and then to a verbal report. The conceptualist surely need not grant any of these assumptions.

My reply to this case also calls into question Block's treatment of the Sperling experiments (Block 2007, 487–491). Block concludes that Sperling's findings show that "phenomenology overflows accessibility," meaning that the conscious experience of those tested must have been richer

than their ability to report or access. If the reply in the previous two paragraphs is valid, then it raises serious doubts about Block's conclusion because, for example, either conceptualization of all the letters doesn't happen in the first place or conceptualization of the letters does occur but does not remain in memory for the purpose of verbal reporting. This reply is also augmented by the deflationary strategy explained in section 6.3.2. Furthermore, recent experimental results more directly contradict Block's interpretation of the Sperling results. De Gardelle, Sackur, and Kouider (2009) found that participants persisted in the belief that only letters were present when pseudo-letters were also included in the array. This belief persisted even when participants were made aware that they might be misled. An unwarranted overconfidence persists on the part of participants, challenging the view that there is very rich phenomenology in a brief visual presentation. This supports the position that not all letters are conceptualized initially in the Sperling experiments.

Reply 2: Regarding the dice case, I think that that once Mary has acquired the concept DODECAHEDRON, it is really her *later memory episode* of playing the game, not the original experiences, to which this concept is applied. Thus Mary really later *infers* that she was playing with a dodecahedral die (Speaks 2005, 385). She comes to *believe* or *know* that she played with such a die. But it doesn't follow from this that the content of her original experience contained DODECAHEDRON. Contrary to Martin's claim, if we are to take this case seriously, I would think that it did *not* initially seem to Mary that she was throwing a twelve-faced die. Given Martin's own description of the case, it seems to me that one might find it equally plausible to suppose that Mary's prior experience was instead of a "many-sided figure." Indeed, Martin tells us that Mary really only distinguishes the dice by the different distributions of colored spots on the faces. Mary does not think of the dice *as* distinguished by specific shapes, but rather perceives both dice as many faced. Now, *if* this really makes sense, then it seems to me that she did not apply DODECAHEDRON or OCTAHEDRON at all to the dice in her experience before acquiring these concepts. That is, *if* Mary is really only seeing each dice as many faced, then I think she is experiencing the dice, or the *shapes* of the dice, in the same way.

Perhaps Mary, owing to her deficiencies, is similar to a normal person with respect to, for example, discriminating between a 100-sided and a 99-sided figure. I would think that few, if any, of us could make such discriminations, but yet we surely still have the relevant concepts. This suggests that what is often viewed as a very weak condition on concept possession (discriminatory ability) is perhaps too strong in such cases.

Of course, the main problem becomes that some of Martin's hypothetical scenario is a bit hard to swallow. Would someone so deficient in math and geometry play such a game? Martin describes Mary as a keen player of board games, often using unusual dice. Could she really be so "keen" without having *some* concept of a dodecahedral die? Once we look more closely at the case, it might turn out that Mary must have had the above concepts all along. She presumably does independently have the concepts EIGHT and TWELVE; how could she have completed elementary school without them? Doesn't the game involve those concepts in other ways? Doesn't Mary know that twelve is greater than eight or that eight plus four equals twelve? She also presumably *feels* the dice when she plays the game—don't they feel different enough so that she could at least know how to differentiate them via that sensory modality? Even if she differentiates the dice solely in terms of colored spots *during the game*, doesn't she ever notice at other times that there are more sides to one die than the other? Didn't she ever just count the sides? Mary is obviously not the curious sort, despite her prowess in playing board games. Can Mary differentiate between two shapes when she is, say, seeing two stained-glass windows of different proportions?

In any case, it starts to become extremely unclear that Mary really did not have the concepts in question, according to our criteria in CONPOSS. Moreover, if Mary had the concepts TWELVE and SIDE, it would seem that she would have the concept TWELVE-SIDED. This would be a minimal case of compositionality and perhaps a case where one concept (DODECAHEDRON) is both easily decomposed into two simpler concepts and a rare definitional case of a necessary and sufficient condition. The issue is presumably not simply whether Mary had ever heard *the word* "dodecahedron" but whether she had at least some grasp of the concept.

But *if* Mary really cannot differentiate eight-sided dice from twelve-sided dice at all, and she really cannot recognize or identify the dice as having those properties, then I don't see how she would have a conscious visual experience with a content that includes DODECAHEDRON or OCTAHEDRON. Of course, we have few, if any, details of the game itself that could help to settle these issues. So much the worse for another rather unhelpful thought experiment.

My reply here has some affinity to the so-called speckled hen case, as discussed, for example, by Tye (2009b, 16–18). This appears to be a problem for CON because, for example, it seems that we are conscious of all the speckles without applying the correct number concept. However, first, it is unclear that one would be conscious of *all* the speckles if that means paying attention to *each* speckle separately. This seems unlikely unless perhaps one

goes through a lengthy and detailed examination of the hen, and thus the deflationary strategy is also relevant here. Consciously experiencing a hen with lots of speckles is perhaps more like experiencing the Marilyn Monroe wallpaper except on a smaller scale. Alternatively, it is perhaps more like experiencing the 99-sided and 100-sided figures described above. Second, it is entirely unclear to me how two subjects (or one subject at different times) could have the *very same* perceptual experiences while applying, say, different number concepts to the hen. Conversely, if two subjects are applying the exact same object, color, and number concepts to a speckled hen, then it is not clear to me how they could have different conscious experiences. If, say, the experiences are like change blindness cases and thus cannot initially be differentiated, then I am inclined to hold that they are the same. Tye (2009b, 17), no friend of CON, similarly explains that "one can be conscious of the speckles on the hen without each speckle's being such that one is conscious of *it*." Tye agrees that there are speckles of which one is not conscious. On the other hand, there is also a "collective" sense in which one sees all the speckles at once.

Reply 3: Finally, regarding the mustache case, I am not sure what to make of the claim that "one did not *notice* the moustache during the initial encounter with the General (or Fred)." In what sense is this true? In the General's mustache case, we are told that you are talking to and facing the General for most of the night. How could you do so without *once* noticing his mustache? We are presumably not talking about a case of inattentional blindness. We are also not talking about an obscure or somewhat hidden part of his body or article of clothing. I do not see why we should grant this supposition in the initial hypothetical case, as Chuard also remarks (2007, 39). The problematic assumption, then, is that "if we do not later remember something, then we must have failed to notice it earlier." This does not seem right. For one thing, we have all noticed and conceptualized many facts while studying for a test, only to be disappointed when we cannot remember each and every fact later. Alternatively, I may be extremely attentive and notice all sorts of things about a basketball game that I am watching, but still be unable later to remember or report on most of it. It again seems at least equally plausible to think that we often do not remember what we have previously noticed.

Now, the conceptualist may seem committed to the claim that "only what a subject notices or attends to at time t is represented in an experience at t." This would be based on the view that "only what a subject notices or attends to at time t is conceptualized at t," which seems to be supported to some extent by the deflationary strategy discussed earlier. These claims are,

however, a bit too strong even for a conceptualist. For example, Dretske might object that one can merely "see" things that one does not notice at the time, as we might also say with respect to at least some cases of change blindness, which he prefers to call *difference blindness* (Dretske 2004, 2007). It is not as if we are *entirely* blind to the changed features of a complicated scene after viewing it initially even if we are blind to the *difference* between the scenes. Thus we need to be more precise. Recall first the difference between focal and peripheral awareness. As we will see in the next subsection, we also need to distinguish between applying fine-grained and coarse-grained concepts to objects and properties. With these distinctions in mind, it is more reasonable for the conceptualist to hold that

(A) Whatever is consciously experienced at time t in a fine-grained way within one's focal awareness is conceptualized in a fine-grained way at t, while whatever is conceptualized at t only in a coarse-grained way outside one's focal awareness must be experienced in only a coarse-grained way at t.

But it is again still perfectly possible for a subject S *to see* something, even in one's focal awareness, that is *in fact* an object O with property F, but, for example, conceptualize O in a more coarse-grained way given a limited conceptual repertoire. I return to this theme in the next subsection.

Finally, and more generally, there is something odd about relying so heavily on memory to try to falsify CON. After all, as is well known, even simultaneous phenomenological reporting is often fallible. The memory cases would thus add another layer of doubt or fallibility. Moreover, how does one separate the way an experience might seem later (via memory) once additional concepts are acquired from the way it seemed at the prior time without those concepts? Martin rightly recognizes that "memory is a notorious deceiver" (1992, 242) and that "it cannot be denied that one's later conceptual sophistication often does alter one's memories" (244). But he still insists that his rather unusual hypothetical memory-based case is successful against CON. In any case, there are numerous responses to the above cases that, I think, show that conceptualism is defensible.

Part of the overall problem here is just how to characterize the relationship between attention and consciousness (Mole 2008, 2009). It is natural to suppose a close commonsense connection between attention and consciousness, such as:

(B) Attention is *necessary* for consciousness; that is, consciousness requires attention.

However, as I have made clear especially in chapter 5, consciousness seems to be a broader category than attention due, for example, to the existence of peripheral (conscious) awareness. So it does not seem that attention is *always* necessary for consciousness. (B) is much too strong. Some might urge that inattentional blindness also shows that attention is required for conscious perception. Indeed, Mack and Rock (1998) themselves draw such a conclusion. For example, when one is paying attention to something (passing the basketball) even within one's focal consciousness, many subjects are not conscious of other objects within that awareness (the man in the gorilla suit). Thus no conscious perception of O exists without attention to O, that is, consciousness of O requires awareness of O. But, again, *some* subjects do notice the gorilla, and so (B) is again too strong (Mole 2008, 93–97).

The issue of whether or not attention is sufficient for consciousness is perhaps even more difficult. Consider:

(C) Attention is *sufficient* for consciousness; that is, attention requires consciousness.

There seem to be cases where a subject's attention is attracted by a less vivid stimulus without conscious awareness (Jiang et al. 2006). Subjects are presented with attention-grabbing stimuli (such as erotic photographs) to just one eye, which are shown to be unconsciously processed. The more vivid stimulus presented to the other eye, however, draws conscious attention. So one might conclude that the erotic pictures capture the subject's attention even though the subject is not conscious of them. Thus (C) is also too strong. Attention does not *always* require consciousness. Similar conclusions might be drawn from blindsight cases, where patients often successfully guess at some characteristics of a stimulus that is not consciously seen. It would seem that the blindsighted subject has no conscious perception of an object O but is attending to O in some sense, albeit prompted by a questioner. Nonetheless one could make the case that (C) is perhaps ambiguous between attending to a *location* and attending to an *object*. So even if (C) is too strong, it is perhaps still true to state:

(D) If a person is *attending* to a thing, then the person is *conscious* of that thing.

It seems much harder to find a counterexample to (D). A related problem with (C) and (D) has to do with ambiguities between state and creature consciousness, as well as voluntary and involuntary behavior. We might say that the blindsighter is generally creature conscious but not state conscious of the object in the blind field. Further, if we wish to hold that something

has captured the subject's attention, but without state consciousness, then it is clear that the notion of "attention" at work is some sort of *involuntary* creature consciousness, such as occurs with involuntary eye saccades or unconscious processing (as in the erotic photographs mentioned earlier). In this sense, it is possible to attend to things of which one is not conscious. But it still seems otherwise plausible to suppose that whenever a creature pays *voluntary* attention to a specific stimulus, there will be state consciousness of that stimulus. Finally, the notion that attention is not sufficient for consciousness is sometimes seen as being supported by neurophysiological evidence showing that distinct neural pathways are responsible for attention and consciousness (Koch and Tsuchiya 2006). I will not pursue this issue further here, but the neurophysiological investigation continues to be a source of rich debate.[17]

6.3.4 HOT Theory and the Complexity Objection

Much of the discussion in the previous sections is also relevant to what I have previously called the "complexity objection" to HOT theory (Gennaro 1996, 89–91). Once again, the main premise is that a subject can be conscious of numerous aspects of any given perceptual experience. Just as the alleged richness argument is used against CON, so there would also seem to be a problem for HOT theory, namely, the level of complexity that HOTs would need to account for the alleged richness of conscious experiences. It would seem that some HOTs must be absurdly complex and thus contain an incredibly large number of concepts. Byrne calls this kind of objection to HOT theory "the problem of the unthinkable thought" (1997, 117). Several responses cumulatively take the sting out of the objection.

First, a HOT theorist might very well use the deflationary strategy described in section 6.3.2. The phenomena of inattentional blindness, change blindness, peripheral awareness, filling-in, and so on show that our conscious states are not as rich as it might seem. Rosenthal also briefly alludes to these phenomena in various places (e.g., 2002a, 416, 421n44). Thus the HOTs in question need not be as complex as it might seem.

Second, even opponents of conceptualism and supporters of the richness argument typically concede that thoughts *do* have entirely conceptual contents. So if complexity is a problem, then it might also be a problem for any theory of mind that allows us to have very complex thoughts.

Third, it might be that there is more than one HOT in some cases of complex experiences. For example, one may have a conscious state that combines the visual image of the waves on a beach, the feel of the water, and the sound of the waves crashing. In at least these multisensory cases,

it seems wiser to suppose that there are several HOTs, perhaps one for each sensory modality, which are bound together to produce a very complex experience. As Byrne points out, although this leads to further questions about the unity of consciousness, one might use this strategy to show that these HOTs are "small enough to be individually thinkable" (1997, 118).

Keeping in mind the richness argument, Rosenthal offers the following related response: "We may need fewer [HOTs] than [it] might at first appear. The content of HOTs may typically be reasonably specific for mental states that are near our focus of attention. But it is unlikely that this is so for our more peripheral states. . . . The degree of detail we are [aware] of in our visual sensations decreases surprisingly rapidly as sensations get farther from the center of our visual field. . . . The content of one's HOTs becomes correspondingly less specific" (1997, 742).

Finally, we must also keep in mind that the HOTs in question are not themselves conscious when one is having a first-order perceptual experience. Thus, if part of the worry is that we cannot hold so much information simultaneously in our conscious minds, then this is not really a problem. Rosenthal explains: "The worry about positing too many [HOTs] comes from thinking that these thoughts would fill up our conscious capacity. . . . We would have no room in consciousness. . . . But this is a real worry only on the assumption that all thoughts are simultaneously conscious thoughts" (Rosenthal 1993b, 209; cf. 2005, 62).

It seems to me that we are also now able to answer Chuard's challenge that conceptualists need to offer a more "positive account of the exercise of conceptual capacities in experience" (2007, 41). Given the structure of the HOT model and the connections between HOT theory and conceptualism described especially in section 6.2, we have seen how such an account goes. It must be remembered that the concepts in question are not consciously deployed during first-order conscious experiences, which fits well with the phenomenological facts. Nonetheless the concepts in the HOTs reflect the content of the conscious state. We can also provide a plausible account of what it is to possess a concept (i.e., CONPOSS) to fill out the story.

6.4 Fineness of Grain

6.4.1 The Main Argument and Initial Reply

Another alleged problem for conceptualism has to do with the so-called fineness of grain in our experience. Thus it is often said that conscious perceptual experience, whether it is rich or not, is much more fine grained than the concepts one possesses. In other words, it seems that one can

experience many objects or properties without having the concept of that specific object or property. For example, a subject could experience a shade of red or a shape without having the corresponding concept, and the same can presumably be said for auditory and other experiences. Thus one seems to be able to have conscious experiences with nonconceptual content, that is, perceptual states that represent the world in ways that do not reflect the concepts possessed by a subject. It would seem that I can perceptually discriminate and experience many more colors and shapes than I have concepts for. A number of authors have raised this argument against CON (e.g., Evans 1982; Peacocke 1992, 2001a; Tye 1995; Heck 2000). One oft-quoted (rhetorical) question is: "Do we really understand the proposal that we have as many color concepts as there are shades of color that we can sensibly discriminate?" (Evans 1982, 229).

The main initial conceptualist reply comes from McDowell (1994), who argues that we can form *demonstrative concepts*, such as "that shade of red" or "is colored thus" for each specific shade of color that is experienced (cf. Brewer 1999, 170–174). We therefore do in fact have enough fine-grained concepts to account for the perceptions in question, though they are special demonstrative concepts rather than the more frequently used "general" or "sortal" variety. We might call this the "demonstrative strategy" (Chuard 2006).

The demonstrative strategy has led to a number of important counter-replies. Perhaps most prominent is Kelly's argument that demonstrative concepts do not really deserve the name "concept" at all, since the subject is unable to *reidentify* things that fall under it: "The re-identification condition states that in order to possess a demonstrative concept for x, a subject must be able consistently to reidentify a given object or property as falling under the concept" (Kelly 2001a, 403). Kelly then uses the example of having many paint chips of different shades of a color, such as red. Now one will typically be able to discriminate one shade of red from others when looking at them simultaneously. Suppose you have picked out red_{27} as the shade you will purchase. But then suppose that you drop the paint chips all over the floor. You would very likely not be able to recognize and pick up red_{27}, that is, identify or recognize red_{27} independently of the others, not to mention recognizing a lone shade of red as red_{27} the next day. If a subject can discriminate red_{27} from red_{28} in perception but cannot reidentify them separately, then we seem to have an example where perception is more fine grained than the concepts one possesses.

This reidentification condition on concept possession seems reasonable at first glance. The idea is that genuine concept possession requires having

some kind of cognitive distance from the context in which the concept had its initial application. Kelly's paint chip scenario also gains support from some empirical literature showing that our *memory* for these fine discriminations is extremely limited (Raffman 1995; Dokic and Pacherie 2001). Indeed, even McDowell seems to support some kind of reidentification condition for demonstrative concepts, even if only for a short time interval. In speaking of demonstrative concepts, he tells us that "what ensures that it is a concept . . . is that the associated capacity can persist into the future, if only for a short time, and that, having persisted, it can be used also in thoughts about what is by then the past, if only the recent past. What is in play here is a recognitional capacity, possibly quite short-lived, that sets in with the experience" (McDowell 1994, 57).

6.4.2 Reexamining the Reidentification Condition

I am not as enamored with the demonstrative strategy as McDowell and Brewer are. Nonetheless a number of other compelling avenues are available to the conceptualist, some of which have already been discussed in the literature:

(1) Speaks (2005, 380–381) suggests that Kelly's paint chip example is better understood as a *reductio* of the reidentification constraint rather than a problem for the conceptualist. He rightly wonders why, if the choice is between crediting a subject S with having a demonstrative concept (and thus a thought) about a visually present shade of color (even without an ability to reidentify it later), and denying that S has a demonstrative thought at all, one would opt for the latter. Why not just give up the reidentification condition?

(2) Chuard (2006) argues at great length that many different versions of the reidentification condition are not required for having demonstrative concepts and need not be accepted by the conceptualist. To use just one hypothetical example, Chuard points out that if several mysterious-looking stones from outer space suddenly landed on my desk, I would surely be able to have demonstrative thoughts about each of them, but I would not thereby be expected to reidentify them individually if they were moved around my office or be able to identify which stone appeared first. Even if more and more stones start appearing, this does not seem to retroactively undermine the fact that I had earlier thought demonstratively about each stone.

(3) Brewer (2005) accepts the reidentification condition but argues that in the paint chip case, an *appropriate* reidentification condition *can* be met. This is because demonstrative concepts are indexed to the particular occasion of discriminating between two shades of color. As Kelly recognizes

elsewhere (2001b), demonstratives are context or situation dependent. After all, isn't this the point of using a demonstrative? Thus Brewer contends that it is a mistake to require that S be capable of recognizing or reidentifying the correct paint chip shade after a complete break in experience. He explains that the most a conceptualist should require is that S be able to *keep track* of the object or property over time or during changing viewing conditions.

Brewer thus suggests that the initial discrimination is irreducibly *relational* in content, such as of the form "colored-thus-in-relation-to-that." This more complex demonstrative color concept is essentially context dependent. Isn't that what demonstratives are supposed to be? And *this* demonstrative concept can satisfy the reidentification condition, that is, S surely *can* discriminate the two samples again at some later time. Kelly (2001a, 415–416) considers a similar reply but dismisses it much too quickly. Brewer's alternative of using relational demonstratives seems like a plausible way to handle such cases. The idea is that during the initial discrimination of the two samples, there is but one perception of a difference between two shades rather than two different perceptions.

(4) Following on this theme, I suggest that we often deploy what might be called *comparative* concepts, such as in thoughts of the form "that shade is *darker* than this one" or "this object looks *bigger* than that one" or "that note sounds *lower* than this one" and so on. The idea is that some perceptions are fundamentally characterized as *comparisons* (or contrasts) between two properties or objects. Rosenthal sometimes also speaks of comparative concepts (2005, 188–189, 204–207), though he is not explicitly concerned with defending conceptualism. He says, for example, that "two discriminable shades of red may differ because one is slightly darker . . . slightly brighter, or slightly more like the color of some particular type of object. We have a huge range of comparative concepts available for HOTs to use" (2005, 188). He further explains that "even when seeing two very similar colors together results in two plainly different color experiences . . . it may well be that seeing each separately would result in color experiences that are subjectively indistinguishable. . . . When we see the two together, our conscious experiences of them differ because of the comparative concepts that figure in our HOTs. When we see the very same colors separately, there is seldom any way to apply the relevant comparative concepts; so the resulting conscious experiences are subjectively the same" (2005, 189). So, for Rosenthal, the comparative concept in a HOT involved in experiencing two shades of red is absent when only one shade is presented to a subject. This seems right to me.

It is also crucial to note that for the conceptualist, what matters most is *the way* that the subject experiences these very similar properties or objects. And, as I have argued, the concepts one possesses determine the content in question. It may very well be that a nonexpert on colors will have different color experiences from a painter or artist who may be able to reidentify each color later. And it is important to recognize that the demonstrative "that shade" refers to the way that one experiences the shade in question (or the shade-as-experienced), not to the shade that is experienced. As Kelly recognizes, the same color shade could be experienced differently depending on changing lighting or environmental conditions. Stated slightly differently, a shade that is *in fact* red_{27} might not be experienced as red_{27} for a variety of reasons, including the conceptual repertoire available to the experiencing subject. I return to this theme in the next subsection.

(5) It is not even clear to me that reidentification is always necessary for having some degree of a *general* concept, let alone a demonstrative one. It is important to recall that concept possession comes in degrees. Most of us surely have the concepts ALLIGATOR and CROCODILE, at least to some degree, but it might take some work to learn how to distinguish one from the other reliably. As it turns out, one main difference is that alligators have wider and shorter heads than do crocodiles.[18] Of course, many who have at least *some* concept ALLIGATOR or CROCODILE will not know this. Suppose, however, that a subject S *is able* to reliably discriminate alligator pictures from crocodile pictures. Now suppose that S is given only one picture and asked to identify it as one or the other. It is unclear that S would be able to do so, or at least to do so as consistently as Kelly demands. The head size comparison is no longer visually present to S. The same could go for any number of very similar animals, say, various kinds of snakes or birds. It still seems wrong, however, to suppose that S therefore has *no* CROCODILE or ALLIGATOR concept on this basis; after all, S would likely know many central features of each, such as they are reptiles, four-legged, typically found near or in water, large teeth, and so on. Indeed, the reverse scenario might also be true; that is, one might correctly identify alligator pictures individually but then have difficulty when pressed to choose between two pictures, one for each animal.

Here is another example: With some brief training, one can become aware of some differences between the sounds of a piccolo and a flute. One main difference is that the piccolo is capable of making sounds of a higher pitch than is the flute. The pitch is, or can be, different. So I might be able to discriminate piccolo sounds from flute sounds, especially if they are presented simultaneously or nearly simultaneously. Nonetheless, if I hear sounds from just one of them an hour later, it is not clear that I will

be able to reidentify that sound. Yet it seems to me that I still have at least some concept of PICCOLO and FLUTE, and even the SOUND OF PICCOLO and SOUND OF FLUTE.[19] The point again here is that we can discriminate between two sounds (or color shades, and so on) using comparative concepts at one time while not being able to recognize or reidentify only one of those sounds (or color shades, and so on) at a later time.

Recall also the earlier case of the multisided dice. Surely there is some point at which all of us can quickly discriminate between an n-sided figure and, say, an n + 2-sided figure at a given time, but not be able to reidentify the n-sided figure alone at a later time. For example, I may be able to discriminate between a 20-sided and an 18-sided figure but have difficulty reidentifying only the 18-sided figure later.

In any event, these cases are much like the examples of the shades of red and the paint chips. These kinds of fine-grained cases should also cause us to acknowledge that we must interpret CONPOSS carefully with respect to very similar properties or objects and thus when extremely fine-grained concepts are involved. The phrase "to some extent" in clauses (a) and (b) is extremely important, and it may just be that, in some cases, we should only demand that a subject be able *either* to have discriminatory ability *or* to have a recognitional (or identification) capacity. We should not demand that a person have superhuman abilities to possess a concept.

Recall that a biologist or musician will likely have different experiences from a nonexpert in some fine-grained cases, as we have seen with the familiar wine-tasting example. And an experienced painter, for example, may even be able to reidentify red_{27} later. Experts' additional conceptual knowledge provides them with a superior degree of concept possession, which, in turn, affects their perceptual experience and ability to identify instances of that concept. But as long as the experiences are dictated by one's conceptual repertoire, including comparative concepts, conceptualism is left unthreatened.

So in the two paint chips case, we might acquire more specific conceptual content at the lower-order level (such as RED_{26} and RED_{27}), which is more detailed and fine grained than those concepts present at the HOT level, at least initially. Any given subject, however, may have only coarse-grained concepts such as DARK RED in the HOTs and will thus have only the more coarse-grained color experience when presented with the two paint chips separately. The crucial point here is that even though there isn't an *exact* match between all the concepts at each level, the HOT's concepts, which contain concepts that S already has, are consistent with, but more general (or coarse grained) than, the LO concepts. Suppose that red_{27} is lighter than

red_{26}. S may not be able to distinguish red_{27} from red_{26} when presented with each separately. However, S would still likely be able to distinguish them when presented together because S may have, say, RED_{27} and DARKER or simply DARK RED and LIGHTER. Much the same is true for the cases of multisided figures, very similar sounds, recognizing similar animals, and so on.

(6) Much as in the earlier discussion of ambiguous figures and associative agnosia, then, we need to modify premise (2) from the argument in section 6.2 that said that "whenever a subject S has a HOT directed at *e*, the content *c* of S's HOT determines the way that S experiences *e* (provided that there is a match with the lower-order state)." I changed it to read:

(2′) "Whenever a subject S has a HOT directed at *e*, the content *c* of S's HOT determines the way that S experiences *e* (for whatever *full or partial* conceptual match exists with the lower-order state)."

However, due to some of the examples addressed in this chapter, it seems that the situation can further be complicated in at least three ways: (a) If the HO state has more specific conceptual content than the LO state, then we have a case where the conscious perception is more fine grained than if one did not possess those concepts because the HO concepts recognize the LO concepts in a more specific way. Examples of this would be ambiguous figures and associative agnosia. So, for example, associative agnosics would seem unable to apply the more general concept (such as WHISTLE) at the HO level. More specific concepts at the HO level, such as SILVER, are what determine the way that the conscious state seems to the subject. Alternatively, when one perceives an ambiguous figure (such as the vase–two faces), the LO input is itself so general that the subject can apply more than one specific concept to it at different times. (b) If we suppose that a LO state can be caused to have more specific conceptual content than the subject can already incorporate into a HOT or MET, then the LO state contains more fine-grained concepts than the HO state has. In this case, one's conscious perception will be more coarse grained than if one already possessed the more specific concepts because the HOT (or MET) can only recognize the LO content to that extent. An example of this might be when one initially sees a more fine-grained color than one has the concept for. (c) However, there is also the intriguing possibility that, instead of the previous explanation, a more specific concept in the LO state can still be "matched" by the HO state because of a *combination of comparative and existing HO concepts*. For example, RED_{27} in the LO state might be experienced because the HO state has the concepts RED_{26} and LIGHTER. Thus we should revise premise (2′) to read:

(2″) "Whenever a subject S has a HOT directed at *e*, the content *c* of S's HOT determines the way that S experiences *e* (provided that there is a full or partial conceptual match with the lower-order state, or *when the HO state contains more specific or fine-grained concepts than the LO state has*, or *when the LO state contains more specific or fine-grained concepts than the HO state has*, or *when the HO concepts can combine to match the LO concept*)."

6.4.3 The Priority Argument

Despite the importance of the demonstrative strategy and the replies in the previous section, it seems to me that there is a much deeper and more interesting issue in the background. It ultimately has to do with concept acquisition, which I address at much greater length in the next chapter. But let us set the stage here in this context.

What I have in mind is what Bermúdez and Cahen (2010) call the *priority argument*. Even if some version of the demonstrative strategy is defensible, the problem remains that demonstrative concepts do not seem to be *explanatorily basic*. That is, they don't seem to explain how, say, I experience a completely novel shade of red (say, red_{17}). Indeed, this seems explanatorily backward. It looks instead as if the experience of red_{17} must be *prior to* the possession of the demonstrative concept. Having the experience allows me to form the demonstrative concept in question. But, the objection goes, the conceptualist presumably holds that what explains the (content of the) visual experience are concepts that the subject *already has*. Indeed, one common objection to conceptualism (and to HOT theory, for that matter) is that it cannot explain concept acquisition generally (Speaks 2005, 368–369; Peacocke 2001a, 252–253; Roskies 2008). How can I acquire a concept if it (or something very similar) is already presupposed in experience?

This point has not been lost on some authors, though it has not generated the attention it deserves. Perhaps most prominently, Heck (2000, 492) tells us that "what *explains* my having these [demonstrative] concepts is my having (had) an experience with a certain sort of content. But, if that is right, it is hard to see how these demonstrative concepts could be part of the content of my experience. . . . There would not seem to be sufficient distance between my having the experience and my possessing the concept for the former to *explain* the latter."

In a similar vein, Wayne Wright (2003, 52) remarks that the "possession of the particular demonstrative concept deployed when having an experience which otherwise extends beyond one's conceptual resources does not look to be antecedent to the experience itself. Instead, it seems that the

ability to form the demonstrative in question depends on having already been presented with a suitable sample in experience."

Coliva (2003, 64) explains that in the case of a newly experienced shade of red, the concepts "must *already* be in place in the discrimination of the shade in question," and so it is unclear how forming demonstrative concepts helps the conceptualist. Finally, A. D. Smith (2002, 113) uses the example of the shade of red called "vermilion" to make a similar point. He points out, in typical empiricist fashion, that I can come to acquire the concept VERMILION, but only because I can now *already* see vermilion things.

The priority argument is perhaps the most difficult problem facing conceptualism. There are, however, at least two responses:

(1) One move would be to deny that the experience *causes* the concepts to form, and so the experience is not temporally prior to the concept. Although Heck flirts with this reply (2000, 492–493), it is explicitly defended by Brewer (2005). In essence, Brewer argues that the conceptualist might claim that the experience of a color sample, red_{17}, *just is* the entertaining of a content within which the demonstrative concept THAT SHADE is a constituent. The idea is that the way that experiences *explain* demonstrative concept possession is *constitutive* rather than causal. Thus a subject deploying a demonstrative concept RED_{17} also explains the experience in the sense that the experience is constituted by the concept.

(2) More interesting, however, is the implication that the priority argument has with respect to concept acquisition generally, including for very specific fine-grained experiences. We have already seen, for example with associative agnosia, that one can "see" or "look at" object O but not recognize it *as* an O. We have also seen that *comparative* concepts are often applied to a given experience. Thus, with regard to experiencing an entirely novel shade of color red_{17}, I suggest that a subject S *need not* see it initially *as* red_{17} but see it in whatever way S's concepts will allow at that time. The conceptualist could maintain that S, for example, is *initially* experiencing red_{17} simply *as* a shade of red that is darker or lighter than other shades that S has already seen. S already possesses SHADE, DARKER, RED, LIGHTER, and so on.

If I am walking down the street and see someone wearing a shirt that is a shade of red that I have never seen before, then the conceptualist could hold that, *at least for the initial very brief encounter*, I did *not* consciously experience the shirt *as* that shade of red. Of course, I can quickly *acquire* that concept, which would allow me to see the shirt *as* red_{17} from that point on. In this case, concept acquisition of RED_{17} can occur extremely quickly and *unconsciously*, even just in a matter of a split second. This line of reasoning also allows us to dispense with Brewer's "constitutive" reply and to sidestep

the priority argument regarding demonstrative concepts. This is because we can concede that the experience *of* a novel shade of red *is prior to* the possession of that demonstrative concept, after all. Smith's example of vermilion is instructive here, too. He seems to concede this point when he says that "perhaps, at the [initial] moment, I do not see vermilion things *as* vermilion," and "I am . . . simply not attuned to that particular shade of red" until after I acquire the concept (A. D. Smith 2002, 113).

But in the case of the two paint chips, we could also suppose that a comparative concept is initially applied to that experience and to those two shades of a color. Even if I have never previously seen those two shades of red, I surely *at that time* still do have more *coarse-grained* concepts. Much the same can be said for general concepts, such as when I hear a piccolo or see a crocodile for the first time, or when I first taste a certain brand of wine. This solution might sound odd at first, especially for color experiences, but the conceptualist need only suppose that subsequent experiences of that shade of red are just slightly different from the initial one. But that seems exactly right, as seems clear from Smith's remark about vermilion and the idea that the more expert one gets, the more fine grained one's experiences become. Doesn't a painter or artist experience some colors differently than nonexperts do? Doesn't the biologist experience a crocodile a bit differently than does a child? It would seem so. Once again, additional conceptual knowledge provides subjects with a superior degree of concept possession, which, in turn, affects their perceptual experiences.

Peacocke also seems to concede this point with respect to concept learning and, for example, the way one might see a pyramid without having PYRAMID. He acknowledges that "there is such a thing as having an experience of something as being pyramid shaped that does not involve already having the concept of pyramid shaped" (2001a, 252). However, instead of concluding that there is nonconceptual content in such an experience, the conceptualist should say that the way a subject does experience a pyramid for the first time is exhausted by already acquired concepts, such as TRIANGULAR, LARGER THAN, THREE-DIMENSIONAL, and so on. Once again, there is a difference between experiencing something that is *in fact* a pyramid and experiencing something *as* a pyramid. This case of experiencing an object with many properties is more typical than the paint chip case. The paint chip case is a highly atypical scenario, that is, one where only a color property is experienced in isolation. We don't normally experience *only* a color shade or, say, a shape. The case of the pyramid is much more typical and thus allows subjects to build on previously acquired concepts.

Finally, if we couple HOT theory with CON, we find a natural explanation of the account of pre- and postconcept possession sketched above, namely, that when a subject S has a conscious experience of something new at t_1, the concepts already possessed by S at t_1, which constitute a HOT, determine the nature of S's experience at t_1. If S later comes to possess additional concepts or some concepts more fully, then S's later experience at t_2 will be different from the experience at t_1. Of course, I have said very little thus far about how exactly concepts are acquired, but that is the main topic of the next chapter. The point for now is simply that conceptualists should hold that when one acquires, or more fully acquires, a concept C at t_2, one's experiences of objects or properties with respect to that concept are different than they were at some previous time, t_1, when one did not possess C.

In closing, then, I tentatively conclude that the Conceptualism Thesis is true. I say "tentatively" only because it remains to be shown over the next two chapters that the Conceptualism Thesis is consistent with the Infants, Acquisition, and Animals Theses. I do conclude here that the Conceptualism Thesis not only is consistent with HOT Theory but flows naturally from it in some enlightening ways. I also conclude that conceptualism can fend off the fineness-of-grain and richness arguments. I now turn to the Acquisition and Infants Theses.[20]

7 Concept Acquisition and Infant Consciousness

Having established the plausibility of the HOT Thesis and the Conceptualism Thesis, it is time to turn our attention to the Acquisition and Infant Theses. This chapter aims to establish that the vast majority of concepts are acquired and that infants are conscious. In doing so, it addresses the issues of innateness and recent work in developmental psychology. Another goal of this chapter is to show that HOT Theory and conceptualism are consistent with the Acquisition and Infant Theses.

More specifically, after briefly explaining what I call the "real hard problem" (sec. 7.1), I argue against radical nativism about concepts but for the more limited "core nativism" (sec. 7.2). I then offer a theory of early concept acquisition with an emphasis on implicit learning (sec. 7.3). In section 7.4, I show how my account can avoid the charge that conceptualism is inconsistent with concept learning, focusing on rebutting arguments offered by Adina Roskies. In section 7.5, I show that infant consciousness is consistent with HOT theory; that is, infants are capable of having at least primitive forms of the requisite higher-order or metapsychological thoughts.

7.1 The Real Hard Problem

I suggested in chapter 4 that concept acquisition may be the *real* hard problem of consciousness instead of Chalmers's more familiar version. We also saw, at the end of chapter 6, that one problem of concept acquisition can be raised with respect to conceptualism under the name "the priority argument." Concept acquisition seems even more difficult to explain for both the conceptualist and the HOT theorist. Even if the reader accepts all my conclusions thus far, many difficult questions remain: Does one have to be conscious first to acquire concepts (or at least most concepts)? If so, how does coherent conscious experience get started in the first place? Are there innate concepts? If so, what are they? How can we acquire concepts

if conscious experience presupposes having concepts (or conceptualism is true)? How can one acquire concepts if conscious experience itself involves a HOT (with its constitutive concepts)?

7.2 Innateness

7.2.1 Against Radical Nativism

There are some legitimate definitional questions regarding the very notion of innateness (Griffiths 2002; Samuels 2007). It does indeed appear that virtually any alleged defining feature is neither necessary nor sufficient. Nonetheless we might simply understand *innate concepts* as those concepts that we never acquire, that is, concepts that we bring to experience from birth (or even before birth). It is perhaps best simply to construe innate concepts operationally as those concepts whose acquisition cannot be explained or do not seem to be acquired at all. This is akin to Chomsky's "poverty-of-the-stimulus" test, which we can apply to concepts as the claim that if nothing in an organism's past experience could explain how that organism acquired a concept, then the concept is innate (Prinz 2002, 193). We must also keep in mind our criteria of concept possession in CONPOSS throughout this chapter.

One might opt for a related account that has come to be known as "primitivism" (Cowie 1999; Samuels 2002, 2007). This is the view that innate psychological "traits" are those whose presence cannot be explained using the explanatory resources of psychology. This account has serious problems, such as apparently counting as innate a trait when or if it is acquired by whacking someone on the head or via brain damage (Samuels 2007; Khalidi 2007). My main point here, however, is to distinguish between concepts that can be explained with reference to *conscious* psychological activity as opposed to some other kind of *unconscious* acquisition, yet still with reference to psychology or cognitive processes. The importance of this will become much clearer later in the chapter.

It is also crucial not to use a weakened notion of innateness to the point where nativism becomes trivially true. For example, Locke long ago recognized that innate ideas or "principles" cannot merely be innate *capacities*, for otherwise *all* ideas would be innate, since we have the capacity from birth to learn whatever we know now. It is also necessary to distinguish innate *structures* or *mechanisms* from innate *concepts*. While the former could be the same as the latter, an innate structure or mechanism is a much broader notion that can include, for example, genetic and biological traits that are entirely nonconceptual and even nonmental. These traits clearly

do not fall under the domain of CONPOSS, whereas concepts are particular mental representations.[1] At the same time, however, we need not suppose that innate concepts are automatically manifested at birth, since some mental representations might not be activated until a later time, such as in the first year of life. As we will see, the key to identifying innate concepts has more to do with identifying concepts that do not seem to be learned or acquired in any sense.

Jerry Fodor (1981, 1990, 1991) is perhaps best known for advocating the view often called *radical nativism*, according to which virtually all lexical concepts are innate, though he seems to have changed his mind in Fodor 1998. Lexical concepts are concepts that can be expressed using a single word, as opposed to complex concepts that are composed of other, more primitive concepts. In addition, according to Fodor, most concepts are primitive and unlearned. Following Laurence and Margolis (2002), we might state his main argument as follows:

(1) Learning requires hypothesis testing.
(2) Hypothesis testing requires conceptual structure.
(3) Lexical concepts are not conceptually structured.
Therefore,
(4) Lexical concepts are not learned.
Therefore,
(5) Lexical concepts are innate.

Thus Fodor's radical nativism goes hand in hand with an atomistic theory of concepts according to which lexical concepts do not have constituent structure and thus are not built up out of more primitive concepts. Only complex concepts can be learned. Fodor is also concerned to show that no other theory of concepts can account for the essential compositionality of concepts. For example, he argues that many complex concepts do not have prototypes, such as GRANDMOTHERS WHOSE GRANDCHILDREN ARE MARRIED TO DENTISTS. Similar arguments are used against other theories of concepts. Recall from chapters 2 and 6, however, that we need not be wedded to the view that a single account of conceptual structure can explain every aspect of concept possession or acquisition. One might adopt a conceptual pluralism while remaining sympathetic to a causal theory of content.

There are a number of standard replies to the foregoing line of argument.

First, it is unclear that learning, or at least *all* learning, requires hypothesis testing, as is stated in premise (1). Learning is often a personal-level and rational phenomenon, but there seem to be other possibilities such as nonrational causal explanations involving internal mechanisms that reliably

correlate to features of the environment. Some concepts also seem to be learned on the basis of a single learning episode or trial. At the least, I think it is a mistake to rule out infant learning virtually by definition. It seems unlikely that infants engage in hypothesis testing.

Second, one might still insist that "learning" is, by definition, a complex, conscious, and rational process. But then we ought to make the further distinction between concept *learning* and concept *acquisition*. After all, given this distinction, the inference from unlearned to innate is clearly fallacious, because a concept might still be *acquired* without involving what Fodor calls "learning." Many authors have raised this point. For example, Sterelny explains:

> Fodor seems to move straight from the idea that primitive concepts are unlearnt to the idea that they are innate. But this inference surely isn't sound. . . . Fodor has at most shown that *certain kinds* of learning don't constitute the acquisition of primitive concepts. It doesn't follow that the concepts are innate; they may be acquired by a different kind of learning process, or acquired but unlearnt. Perhaps not all *acquisition* is learning. (Sterelny 1989, 126; cf. Samet 1986, 580)

In the context of discussing Kant's "Categories" in contrast to the views of both Locke and Leibniz, Bennett (1966, 95–99) also anticipates this point:

> Within the genus concept-acquisition there is the species concept-learning. . . . Nothing is logically prerequisite to a concept's having been acquired except its being not possessed and then later possessed; but . . . [learning] involves the active, rational co-operation of the learner. . . . [Innate] means "possessed but not acquired." (97–98)

In any case, one problem is that Fodor's argument starts with an unnecessarily sophisticated or restrictive notion of "learning." Fodor does ultimately recognize that being unlearned does not entail innateness, but he argues that extant theories of concepts still cannot explain how lexical concepts are acquired. Fodor does recognize that there is the alternative of "brute causal" triggering of mental representations, but insists that this is not a *psychological* process, and so the concepts in question are still innate in an important sense.

Third, premise (3) is also independently controversial because of its embrace of informational atomism. For one thing, one might argue that at least some lexical concepts (e.g., UNICORN) are acquired via the *imagination* to construct what Locke called "complex ideas," sometimes with the help of *reflection*. This would involve something like *compositionality*, and thus conceptual structure, without hypothesis testing. Moreover, there is also the problem of how to handle so-called empty concepts, such as ROUND SQUARE. It would seem, for example, that these empty concepts would absurdly be

treated as identical or having the same meaning for Fodor because they have the same extensions.

Fourth, radical concept nativism is called "radical" for a good reason. The idea that concepts like UMBRELLA, DOORKNOB, and CAT are innate surely flies in the face of common sense, for whatever that is worth. Similarly, it is odd to suppose that nativism could account for the entire range and apparent unboundedness of concepts that we could possess.

To my mind, then, the above four points have the cumulative effect of rendering the rather extreme radical nativist thesis false or, at least, extremely undesirable. Of course, we still need a more positive account of concept acquisition, which is the real difficult chore and will come later in the chapter. It is precisely this problem that led Fodor to embrace radical nativism in the first place.[2]

7.2.2 Core Nativism

Despite my rejection of radical nativism, I think that it is plausible to suppose that there are at least a handful of innate concepts, or at least concepts that we possess from extremely early infancy. Even those with strong empiricist leanings might adopt a limited form of nativism. A number of authors have developed such a view, most prominently developmental psychologists such as Spelke (1998), Baillargeon (1987, 2008), and Carey (2009).

We might call this *core nativism*, according to which a core number of innate concepts are needed initially to "get conscious experience started." These concepts are precisely those which are the most difficult to explain via concept acquisition. They also seem to be present early in infancy and are immediately applied to perceptual experience. So which concepts belong to this elite group? And what is the evidence for their existence? Some authors have, intentionally or not, taken a cue from the list of Kantian Categories, which contain notions such as SUBSTANCE and NUMBER, as well as what Kant called the "two pure forms of intuition," that is, SPACE and TIME. Spelke often refers to "core knowledge" (2007), which, at the least, should involve certain innate concepts.[3]

A number of testing methods are employed to determine which concepts are present in infancy or very early childhood.[4] I will mention just a few.

(1) One is the *familiarization-test procedure* (or the infant habituation paradigm), which examines infants' response to novel items after they are shown a number of objects (or pictures of objects) from the same (or very similar) category. The duration of the infants' looking times at novel items is recorded. The idea is that the longer an infant looks at novel objects (or

pictures), the more reasonable it is to infer that the infant thinks of the comparative objects as belonging to different categories. At the least, there seems to be evidence of category discrimination.

As a general rule, infants will look longer at objects and events that are new. Through the process of *stimulus habituation*, an infant will decrease its looking time when looking over and over again at a familiar object. After a process of habituation, a process of *dishabituation* follows. Once again, it would seem that an infant would look longer at the newer object only if the infant takes it as falling under a different concept. Although this may sound odd to some philosophers, this method has been standard practice in developmental psychology for some time. To be sure, however, psychologists and philosophers disagree to some extent about just how to interpret the results. Nonetheless control experiments are performed to eliminate various other possible interpretations.

(2) The *violation of expectation* method is also used, often in conjunction with (1). An infant's looking times are taken as a measure of whether the infant's expectation was violated by the display. Such expectations seem to reveal concept possession. Infants and adults tend to look longer at surprising and thus unexpected events or outcomes. We look more attentively at objects when something happens contrary to expectations.

(3) The *visual preference paradigm* measures total looking time for two presented objects. Infants generally look preferentially at things that are novel compared to things that are familiar. One finding is that infants tend to prefer patterned surfaces to uniform surfaces and even complex patterns (such as checkerboards with many squares) to simple patterns. These techniques are sometimes used to examine infant visual acuity and pattern perception.

There is a tremendously large developmental literature mainly aimed at discovering just what concepts infants and young children have and can acquire (Murphy 2002; Rakison and Oakes 2003; Gelman 2003; Mandler 2004; Carey 2009). This is a rich and fascinating area of ongoing research, with psychologists often using the term "categories" in addition to "concepts." To categorize is basically the "ability to group discriminable properties, objects or events into classes," whereas a concept is the "mental representation that encapsulates the commonalities and structure that exist within categories" (Rakison and Oakes 2003, 1). Thus psychologists tend to use the term "concept" in the way we have thus far. A *category*, however, has more to do with the *class* of objects or properties that are grouped together, much like what a philosopher would call the "extension" of a concept.

So let us look at the leading candidates for innate concepts. I do not claim that the list is exhaustive or entirely uncontroversial, but to my mind, there is increasing evidence supporting core nativism. As we will see, the list contains many interrelated concepts.[5]

(1) OBJECT/SUBSTANCE. At the earliest stages, infants are able to differentiate (or "individuate") one object from another and track the same object over time. Thus a rudimentary form of object concept seems to be present early on; primitive thoughts can contain OBJECT or THING. For example, four-month-olds already have certain expectations about the solidity and normal movements of objects. In one experiment, infants are shown a solid bar moving back and forth behind another object (an "occluder"), followed with two test trials, one where the occluder is removed to reveal a single bar and another where it reveals two separate bars. Infants stare much longer at the second display, illustrating their surprise that the initial object did not move together as a continuous whole (Kellman and Spelke 1983).[6]

For these and other reasons, Spelke (1990) has argued that infants understand that objects are things that are "bounded, coherent, three-dimensional, and move as a whole." Very young infants use *spatiotemporal* information in establishing object identity (Aguiar and Baillargeon 1999), which is much earlier than Piaget (1954) had assumed. Being able to discriminate and identify objects is of course a key criterion for concept possession in CONPOSS. It seems clear, using the violation-of-expectations method, for example, that infants have at least some rudimentary sense of what counts as a THING. Recall that each of the three conditions in CONPOSS allows for degrees of concept possession. We might say that there is a very *general sortal* concept OBJECT that can be used to individuate and track objects. A sortal concept is a concept that we use to identify and count things in the world. In addition to OBJECT, there are numerous more specific *kind sortal* concepts that are acquired later, such as DOG, TREE, and CAR, as well as *property* concepts (e.g., RED, ROUND). Some have argued that the general sortal OBJECT is used before kind sortals or property information. This is what has been called the "object-first hypothesis," which remains somewhat controversial.[7]

To be more specific about the general sortal OBJECT, Spelke describes several principles of core knowledge, including the principle of *continuity* (objects move on connected, unobstructed paths), the principle of *cohesion* (objects move as connected and bounded wholes), the principle of *contact* (objects do not interact at a distance; that is, only surfaces that are in contact can move together), and the principle of *solidity* (one solid object cannot pass through, or fuse with, another solid object). All these principles

and concepts are interrelated, and many investigators speak of object *permanence* (Baillargeon 1987), and others refer to object *unity*. Baillargeon (2008) has more recently argued that the principles of continuity and cohesion are really corollaries of a more fundamental principle of *persistence*, which says that objects persist, as they are, in time and space.

Many issues arise even at this level, but I will mention just three briefly now. First, it seems to me that if one is individuating one object from another, such individuation must be based to some extent on at least *one* of the object's *properties*. Needham (2001), for example, found that four-month-olds could use texture and orientation, but not color cues, to help them individuate objects. Second, the notion of a "sortal" itself can be extremely problematic and is related to various metaphysical issues of identity. Both philosophers and psychologists recognize this fact (Lowe 2007; Scholl 2007), and a number of literature reviews are available (e.g. Grandy 2008). But for my purposes here, it seems reasonable to suppose that some basic concepts are at work from early in infancy. Third, we must remember that there are degrees of concept possession, and the same is true for core concepts. We need not suppose that infants have a "mastery" of all possessed concepts (Bermúdez 1998, 67), which in turn alleviates some of the motivation for the view that infants have states with nonconceptual content.

(2) SPACE/MOTION/SHAPE/DISTANCE. We have already seen some indirect reference to the notion of SPACE. Infants can track object paths and movement through space. Spelke's principles each contain an element of space and motion. In addition, however, infants seem to have an egocentric perspective that allows them to differentiate shapes and locate objects in space relative to their positions. Concepts such as UP/DOWN, IN/OUT, BELOW/ABOVE, and BETWEEN seem essential to having such ability. It is unclear how an organism could have SPACE, MOTION, and so on without grasping UP/DOWN and so on. At minimum, we have strong evidence that infants grasp a variety of spatial notions in the first year of life (Quinn 2003). Such early spatial cognition would thus also involve rudimentary concepts like the LOCATION of objects and the spatial RELATIONS among objects. Mandler (2004, 2008) places great significance on the idea that many primitive concepts are related to SPACE or spatial relations, such as UNDER, OVER, CONTAINER, MOTION, and CONTACT.

Similarly, infants display ability for *size constancy* as early as eighteen months (e.g., Day and McKenzie 1981). Size constancy is the ability to perceive an object as being the same size despite changes to its distance or orientation. Infants can also represent *shape constancy*; that is, objects maintain their shape through time. This is the ability to perceive an object

as being the same shape despite changes to its orientation or slant. Evidence from newborns supports the idea that infants are able to perceive the objective shape of objects and do not rely solely on the retinal image of those objects (Slater and Morison 1985). This seems to indicate an understanding of relational information. In addition, infants display different anticipatory manual reaches for hidden objects, for example, reaching with both hands for larger objects while reaching with one hand for smaller objects (Rochat 2001, 97–99).

(3) PERSISTENCE/TIME. We have also already seen some reference to the notion of time and persistence in the foregoing discussion. For example, part of the concept OBJECT seems to involve the notion that objects continue to exist through time even when hidden. We might suppose that the more generic notion of DURATION seems present from very early on. Moreover, as Kant argued, it is difficult to make sense of having *coherent* conscious experience unless we presuppose that outer objects persist in time. If we did not make such a presupposition, then we would not be able to distinguish the fleeting subjective succession of our conscious experiences of objects from the enduring nature of the outer objects themselves. Along the same lines, Rochat points out that when infants "understand the notion of permanent objects—that objects continue to exist when they momentarily disappear from view—this understanding is inseparable from the developing sense of the infants' own permanence in the environment" (2001, 79).

In a somewhat Kantian manner, Mandler also treats TIME as closely related to SPACE, the latter being an analogue and extension of the former. She explains that many new concepts are "analogical or metaphorical extension[s] of spatially-based concepts into nonspatial realms" (2008, 222). We might conceptualize time as a "linear path" and think of the "passage of time." We also frequently speak of "time intervals," "time approaching," and the like. Unlike Mandler, however, I would treat TIME as innate and equally fundamental, whereas she seems to take it as being derived from SPACE. As Kant argued, not only is SPACE intimately related to TIME, but we could even think of TIME as more fundamental than SPACE. This is because TIME is the *form* of both inner and outer sense, that is, of both introspection via the temporal spread (or sequence) of mental states and ordinary outer sensory experience of the world via motion and change. Moreover, it would seem that infants understand BEFORE and AFTER (to some degree) which are closely related to TIME.[8] Since it appears that infants have even at least some working memory, BEFORE/AFTER or EARLIER/LATER seems necessary. Indeed, the mere fact that early infants can be tested in the ways described earlier indicates that they grasp the difference between earlier and later moments

in time, can identify the same object over time, and think about different stages of an object.

It is also worth mentioning here that the notion of time is frequently discussed with respect to consciousness and perception. In the case of adults, it would seem that our experience is infused with TIME to make sense of William James's (1890) "specious present" or "temporal extent," together with a combination of working memory and anticipation. For example, it seems that one's experience of hearing a familiar song differs from hearing a novel one owing to the sense of anticipation and knowledge with respect to the familiar song. Our experience becomes richer as we gain knowledge of an *entire* song. We experience *duration* as opposed to some instantaneous moment. Much the same goes for the difference between reading or hearing a foreign language as opposed to a familiar one. We thus understand a conscious experience as representing a continuation of the past and a flowing into the future.[9] These considerations also lend some support to the conceptualist view presented in the previous chapter, that is, with regard to how TIME can enter into perceptual content.

(4) CAUSE. As we have seen with the notion of solidity, infants exhibit clear expectation that physical objects move as a connected whole and do not spontaneously appear or disappear. There is also evidence that infants understand that physical objects are solid; that is, objects do not occupy the same space as other objects. Indeed, the notions of CAUSE and OBJECT/SUBSTANCE are closely related. Some have even argued that "causal continuity" is more fundamental than OBJECT (Rips, Blok, and Newman 2006), though others are not convinced (Rhemtulla and Xu 2007). The idea is that what makes an object endure through time is some sort of deep causal connectedness between the object at one time and at a later time. At the least there often seems to be an implicit concept CAUSE at work when tracking and individuating objects, even if causal continuity is not necessary for object tracking. Moreover, a number of researchers have shown that infants look longer at a noncausal event, presumably because it violates expectations (Leslie 1984). The notion of SOLIDITY is also at work in that infants understand that one (solid) object cannot pass through another (solid) object (Spelke et al. 1992). This is similar to the notion of impenetrability. Finally, the notion of CAUSE also has a clear relation to TIME to the extent that we typically suppose that causes precede their effects in time.[10]

It also appears that adults frequently have conscious perceptions with CAUSE as part of the content. Siegel (2006, 2009), for example, argues for the view that causation is represented in visual experiences in addition to shapes, colors, boundaries, and other more clearly manifest aspects of visual

perception (cf. Bayne 2009; Butterfill 2009). Some support for this view goes back to Michotte's (1963) experiments. For example, adults are inclined to describe scenes of launching and entraining in causal terms, such as collisions. Launching involves cases where object A moves toward a stationary object B and makes contact with B, and B moves in the same direction that A was moving (even though A stops). Entraining is very much the same except that A continues to move together with B. Of course, we need not take such first-person reports at face value, but Siegel presents other convincing cases by contrasting a pair of experiences (e.g., a ball landing in a plant pot just before the lights go out), such that the phenomenal difference between them is best explained by supposing that two events are causally related in one case but not in the other. Much the same seems true for, say, the difference between an experience where one person observes another opening a curtain and letting in some sunlight, as opposed to opening the curtain just as the sun comes out from behind a dark cloud.

Such examples also seem to lend some support for conceptualism. That is, we can see yet another way that CAUSE can enter into the perceptual content of some conscious perceptions. Indeed, these considerations are perhaps best understood as extensions of Spelke's core principles, such as the principle of contact: objects do not interact at a distance; that is, only surfaces that are in contact can move together. At one point, Siegel (2009, 535) considers what she calls the "two-component view," which posits a "higher-order state which represents that the sensory component and the component representing causation are appropriately connected." Siegel rejects this view partly because she claims that such higher-order states would require elaborate self-reflection, which is comparatively rare. However, Siegel does not have HOT theory in mind. I suggest that HOT theory, or at least the WIV, can make sense of the two-component view, since having first-order perceptions only requires an unconscious HOT (or MET), not elaborate self-reflection. Moreover, at least according to the WIV, the two components, including CAUSE in the HOT, are best understood as parts of the identical conscious experience. The so-called sensory component cannot really be severed from the cognitive component. So, unlike Siegel, I take this option to be not an alternative hypothesis but a complementary and supportive explanation.

Another example: I recall once watching part of a football game with someone, AB, who knew very little about the sport. To AB the action was haphazard and chaotic, much as it seemed to my children when they were very young. It seemed to me that our respective perceptual experiences differed dramatically because of the differences in our conceptual knowledge.

Where I saw an organized and well-designed running play with preplanned blocking schemes and the goal of achieving the first down, AB described the action as "guys just running around hitting each other." Indeed, AB could not even see which team had the ball at times, that is, which team was on offense. I saw pass plays as carefully choreographed patterns run by receivers with the offensive line attempting to pick up a blitzing linebacker, AB didn't see that there were various options open to the quarterback and how he avoided the pass rush to find the open receiver. These sorts of examples further indicate how core concepts, such as CAUSE, TIME, and perhaps something more like UNITY, are embedded in one's conscious perceptions. Siegel sometimes speaks of "causal unity" in describing her cases. We organize our experience into coherent, unified, and understandable patterns when we have the requisite conceptual tools.

(5) NUMBER/CARDINALITY. Solid evidence also supports the view that infants have some basic understanding of small numbers (Wynn 1992; Krojgaard 2004; Carey 2009, chap. 4). The idea is that infants can respond to the cardinality of sets of three or fewer. Infants display sensitivity to the cardinality of sets of objects that are moved around and hidden behind a screen. This also serves as evidence for infants' having a sense of "object constancy" as early as three months (see also Baillargeon and DeVos 1991). Moreover, it seems to me that if one acknowledges that infants can track an object over time or individuate one object from another, then there must be some concept of NUMBER involved, at least ONE and TWO. At the least, it seems that ONE, together with a recursive rule (or successor function) for generating the other positive integers through an "add one" operation, is innate (Leslie et al. 2007).

(6) SELF. Some authors have also argued that a primitive notion of "self" must be present even at the earliest stages of infancy (Rochat 2001). I am very sympathetic to this approach. There seems to be at least an extremely primitive kind of "bodily awareness" of self, that is, an ability to differentiate one's own body from other things. This corresponds to what has been called the "ecological self" (Neisser 1991). Thus a rudimentary self-concept or "I-concept" is present very early in infancy. At minimum, an infant has the ability to differentiate between her body and other things, a primitive "me" or "my body" as opposed to "not me."

Using the visual preference method, Rochat (2001, 44–47) also describes how infants as early as three months old discriminate between one display of the child's body from the familiar "ego" or first-person view as opposed to what represents an "observer's" view, looking longer at the latter and unfamiliar view. Newborns also have a bodily sense of self; for example,

they are able to adjust head orientation to the spatial location of sounds. This would seem to indicate "a connection between what they hear and what they intuit about the location of their own bodies in space" (Rochat 2001, 35). Along the lines of the earlier Kantian argument, Rochat also explains that "self-perception is inseparable from our perception of others as onlookers of us" (77).

As we will see, there are degrees of self-concept and self-awareness. An "I-concept" can be extremely primitive in early infancy but becomes more sophisticated as one grows into adulthood. We have already seen something like this, for example, reflected in HOT theory's distinction between unconscious HOTs and introspective states (or conscious HOTs).

(7) *Miscellaneous Concepts*. Finally, it seems to me that a number of miscellaneous concepts are present at birth, including NEGATION (or NOT) and the logical connective IS, which basically involves the concepts EXISTS and AFFIRMATION. These concepts also appear in Kant's list of categories, and with good reason. At the least, they need to be present to allow for the six previous categories. For instance, having the concept SUBSTANCE seems to imply that there IS something that EXISTS. They are also needed for infants to form thoughts with constituent concepts.

It is difficult to see how these concepts could be *acquired from* experience as opposed to *applied to* experience from very early on. Kant, in essence, thought that the empiricists had it backward in claiming that these concepts are derived from experience. Instead, having any experience at all already presupposes having the core concepts just mentioned. For example, how could an infant *acquire* the concept SPACE if her (outer) experience must *already be spatial* from the earliest stage? Similarly, we do not seem to derive NOT from experience, although it is perhaps formed by a combination of OBJECTS and DIFFERENT. If so, it would play an important role very early on with respect to differentiating objects. At the least, we surely do not acquire NOT by experiencing nothing.

Moreover, if my argument is correct thus far, then it seems plausible to suppose that we need demonstratives, such as THIS and THAT, to track objects in space and time, in addition to tracking motion. We might suppose that infants need these concepts to formulate some rudimentary thoughts about objects. Also, implicit in the discussion and methodology here is an ability to employ concepts such as SAME and DIFFERENT. This is particularly clear when infants are tested for object or property individuation. Infants must be capable of having thoughts such as "This object is different from that object," "That object is not the same as that object," "That object is the same as this object," "This body is different from that body," and so on. Indeed,

the most primitive notion of SELF mentioned earlier is basically something like THIS BODY. Other comparative concepts should also be mentioned here, such as LESS/MORE, GREATER/LESSER, and LARGER/SMALLER.

Thus we can see how CONPOSS applies very early on with innate core concepts. Primitive thoughts, which are intentional states, can be constructed from these concepts, which, in turn, also serve to discriminate between objects and properties. Moreover, infants clearly begin to recognize and reidentify objects and properties over time.

In any case, most today reject James's notion that infants merely experience a "blooming buzzing confusion." It seems no longer correct to think of infant experience as an incoherent and confused series of subjective states (Rakison and Oakes 2003, 19). Philosophers have also alluded to many of the foregoing concepts in discussing the nature of infant perception. For example, although not a nativist or a conceptualist, Bermúdez (1998, 2003, 2007b) refers to "canonical object properties," which involve many of the core concepts discussed here. Some of them, such as PERSISTENCE and CAUSE, can be thought of as properties of objects. Noë (2004) observes that visual perception involves a sense of "perceptual presence" within which we take ourselves to experience three-dimensional objects even though a subject S actually sees only the two-dimensional side facing S or when a visual sight line to an object is partially blocked.

One might object that I have committed myself to a form of "theory-theory" of concepts, whereby infants are sometimes viewed as "little scientists" or "physicists" (Gopnik 1996; Gopnik and Meltzoff 1997). This is not the case, however, for several reasons. (1) I have already expressed some sympathy for conceptual pluralism in chapter 2. It is not clear to me that any one theory is correct for all kinds of concepts. Thus even if the theory-theory of concepts is best suited to explain these core concepts, it would not follow that we must endorse it across the board, for example, with respect to concept acquisition. (2) My core nativism is not necessarily committed to the more sophisticated holistic (or functional-role) view of concept meaning whereby the subject has a full-blown theory and is able to make inferences. Rather, having these concepts enables their owner to have what Bermúdez (2007b) calls "perceptual sensitivity" to object features. (3) Recall also that our criteria for concept possession in CONPOSS are fairly modest in the sense that they allow for degrees of concept possession involved in one's discriminatory and recognitional capacity. Thus we can take some insights from theory-theory for a select group of innate concepts but not embrace all aspects of this view.

Of course, even if I am correct thus far, we still have not offered an *account* of concept acquisition in response to Fodor's puzzle of concept *acquisition*, which "amounts to the challenge of explaining how a primitive or unstructured concept can be [acquired]" (Laurence and Margolis 2002). It is important to note that most of the foregoing techniques are mainly aimed at determining which concepts are *already present* in infancy, not at accounting for how an infant *acquires* a concept.

In any case, I now turn more directly to concept acquisition and evidence for the view that lexical concepts can be acquired. This is, in many ways, a more difficult problem that is most central to my purposes.

7.3 Concept Acquisition

7.3.1 Preliminaries

I will focus on the early acquisition of lexical perceptual concepts, since my main interest lies in the overlap between consciousness and concepts. However, I must again first acknowledge that there are many different kinds of concepts, each of which might be acquired somewhat differently. For example, there are abstract concepts such as JUSTICE and BEAUTY, which are likely (consciously) learned at much later stages of development and take more time to learn. There are also sophisticated scientific concepts such as QUARK, which are learned at later stages. As many authors have noted, in grasping the meaning of concepts like QUARK, we are likely to defer to experts in the relevant field and thus have a lower degree of concept possession. Finally, although the classical (or definitional) view of concepts cannot account for most concepts, it may be that we learn some concepts in this way, such as BACHELOR or mathematical concepts like OCTAGON. It may be that language, formal learning, and introspective consciousness are required for the acquisition of these concepts. But there are many other kinds of concepts, the acquisition of which is central to the issue at hand.

In any case, there are two primary distinctions to keep in mind:

(1) So-called *natural-kind* concepts, such as TIGER, TREE, or ANIMAL, as opposed to *artifact* concepts, such as DOOR, DESK, CAR, and UMBRELLA. These are objects that many humans typically encounter early in life.
(2) *Object concepts*, such as HORSE and TABLE, as opposed to *property concepts*, such as RED and ROUND. Of course, object concepts include both natural-kind and artifact concepts. One can readily find in the literature many clever and elaborate taxonomies of concepts, typically in the form of a hierarchical structure (Keil 1989; Chen 2007).

7.3.2 A Theory of Early Concept Acquisition: A First Pass

With regard to object concepts, we might suppose that a perceptual sensitivity is developed over time with respect to objects encountered in experience. Unlike Bermúdez, however, I think that such sensitivity is also conceptual in nature. Once again, key features of concept possession are discrimination and recognition over time, which enable us to perceive objects *as* they are under some description. This typically occurs as a matter of degree upon increased exposure to, and interaction with, the objects in question. Although I focus here primarily on visually presented objects, the same would go for hearing, taste, and so on.

As we saw in chapter 2 with respect to the disjunction problem, a problem arises for what Fodor calls the "doorknob/DOORKNOB problem," that is, how it is that our minds get uniquely "locked onto" their referents. Why and how do we typically acquire DOORKNOB through interaction with doorknobs instead of, say, dogs or tables? Nonetheless, contrary to Fodor, it still seems to me that such concept acquisition is a psychological process, albeit often an unconscious one or at least one that does not require conscious attention directed at the objects or properties in question.

As I explained in chapter 2, I take the most promising approach to explaining representational content, including the content of concepts, to be some kind of causal-informational story. Acquiring a perceptual concept C is a matter of having your mind in a position to be able to respond selectively and reliably to instances of C's. Once again, there is little reason to suppose that this is done in any single way, such as via a definition or prototype. Let's take the concept DOG, for example. Mental states acquire their content by standing in appropriate causal relations to objects and properties in the world. The basic idea is that, say, thoughts containing DOG are about dogs, and *mean* "dog," because dogs cause the thoughts that our minds use to keep track of dogs. And, of course, concepts are mental representations that are constituents of thoughts.

Recall, however, that the disjunction problem shows that a simple causal story cannot properly isolate the correct causal relation. A horse *might* cause the mental tokening of HORSE, but why not SADDLE instead? There is also the related problem of misrepresentation. It would seem that any theory of representation should allow for and explain the possibility of misrepresentation. Perhaps cows sometimes cause the mental representation HORSE. Does HORSE, then, represent *either* cows *or* horses?

In chapter 2, I adverted to the important work of Rupert (1999) and Prinz (2002) as the best fairly recent attempts to handle these problems.

Their work can, at least in part, also be used to explain concept *acquisition* in addition to fixing mental *content*. Recall that Rupert offers a causal-developmental theory according to which there is an actual history requirement for a mental representation to acquire its content accurately. A mental representation R "has as its extension the members of natural kind K if and only if members of K are *more efficient* in their causing of [R] in S than are the members of any other natural kind" (Rupert 1999, 323; italics mine). The notion of "efficiency" was explained in terms of numerical comparisons between the *past relative frequencies* (PRFs) of certain causal interactions. So although every cat is a mammal, the PRF of cats relative to CAT is much higher than mammals relative to that concept. Only PRFs resulting from a substantial number of interactions matter. Thus, in the earlier example, HORSE will not represent cows because HORSE was caused much more frequently by horses.

With respect to *acquiring* TREE, the idea is that TREE is caused by trees if and only if members of the class of trees are *more efficient* in causing TREE in a subject S than are the members of any other natural kind. So, for example, although every tree is a living thing, the PRF of trees relative to TREE is much higher than LIVING THING relative to that concept. Thus HORSE will not also represent cows because HORSE was caused much more frequently by horses (or pictures of horses). The same goes for TIGER, and so on. Notice also that this analysis fits well with the view that concept possession is a matter of degree. S may begin with a somewhat primitive notion of TREE but develop a more sophisticated notion as interaction with trees increases over time. In general, the greater the PRF, the greater the degree of concept possession. There is a lower boundary where the subject would not have any concept TREE, but minimal notions of TREE exist all the way up to some kind of expert level.

Recall that Prinz also urged that the "intentional content of a concept is the class of things to which the object(s) that caused the original creation of that concept belong" (2002, 250). Again, what matters is the actual causal history of a concept. More specifically, mental content is "identified with those things that *actually* caused the first tokenings of a concept (what I call the 'incipient causes'), not what *would* have caused them" (250). So both nomological covariance and incipient causes are necessary to determine intentional content. Much the same, however, can thus be said about concept acquisition. Using Prinz's terminology, then, a perceptual concept C is acquired from encounters with members of the class of c's if (a) C's nomologically covary with encounters with c's *and* (b) c's were the incipient causes of C.

Prinz introduces incipient causes to avoid the disjunction problem facing other closely related theories, such as Dretske's and Fodor's, which are susceptible to the counter that those theories cannot rule out the possibility that (for example) gin could also trigger my WATER concept during the crucial learning period or, that if there were both H_2O and XYZ on Earth, my water concept could still refer to both even if I had only been exposed to one. Of course, as Prinz points out, an already acquired concept can still be modified and strengthened.

This process is psychological because one's mind is learning to detect, and become increasingly perceptually sensitive to, features of one's environment. This process builds up from a primitive base of innate concepts. Concepts, which are mental representations, are always involved in the process even when the process is not itself conscious. One does not have TREE at birth, but one will acquire TREE over time as, say, a more specific instance of a THING with a certain SHAPE in SPACE. Over time a subject will become increasingly perceptually sensitive to trees (or pictures of trees) and will be able to differentiate trees from other things, such as bushes. Eventually one is able to differentiate within the tree category and thus recognize specific kinds of trees. Innate comparative concepts such as SAME, DIFFERENT, LARGER, and SMALLER are also crucially applied in these processes.

With regard to acquiring concepts of natural-kind objects, the process arguably involves something like what Laurence and Margolis (2002) call a "syndrome-based sustaining mechanism" (cf. Margolis 1998). A subject is, over time, able to (at minimum) reliably discriminate between members and nonmembers of a kind without relying on others' assistance. Acquiring the concept CAT, for instance, involves acquiring reliable indicators that something is a cat, such as the shape of a cat, typical motions of a cat, typical sounds of a cat, and so on. These indicators are examples of what I called "central features" in CONPOSS. Thus a "cat-syndrome" is formed and recognition of cats follows. Notice that little prior conceptual apparatus is needed over and above the innate concepts discussed earlier, such as MOTION, SHAPE, THING, and perhaps a few others. This process will also aid a child in differentiating cats from "fakes," that is, objects with the same outward appearance of a natural kind that nonetheless are not instances of the category, such as a stuffed cat or toy dog.

In addition, innate concepts such as IN and OUT are crucial to this stage of conceptual development. One might even make the case that ANIMAL (or ANIMATE THING) is also innate because it involves additional SPATIAL and MOTION primitive concepts, such as the difference between RHYTHMIC (or biological) motion and STRAIGHT (or mechanical) motion. This accounts for the idea that

there is an internal source of motion in animate, as opposed to inanimate, objects.

If, for example, Mandler (2004, 2008) is right, then it is the more global (or "superordinate") notions like ANIMAL that are prior to, say, LAND ANIMAL and to much more specific concepts such as DOG. This suggests that the process of concept acquisition goes from the more general category to the more specific, which echoes much of what we have learned thus far. Innate concepts are extremely general, but they help us to make finer-grained distinctions over time. Concept development thus involves increasing detail, differentiation, and recognition. Experts are those who have the most specific concepts.

What may aid us sometimes in this process is a natural disposition for something like what has been called "psychological essentialism" (Medin and Ortony 1989; Gelman 2003), whereby infants tend to suppose that "hidden" or "nonobvious" properties are taken as a category's primary property (and cause the outer appearances of things). So a typical scenario might run as follows: We acquire CAT in the presence of cats or pictures of cats. Over time, the infant or child recognizes that cats have certain typical observable properties that, however, are likely caused by hidden properties. Recall that CAUSE is innate. The child then *acquires* a certain specific state of mind, which, according to Laurence and Margolis, just *is* a sustaining mechanism. A sustaining mechanism "is a mechanism in virtue of which a concept stands in the mind-world relation that a causal theory of content . . . takes to be constitutive of content. . . . The typical sustaining mechanism is . . . cognitive or inferential" (Laurence and Margolis 2002, 37). It would appear that a sustaining mechanism involves both some memory and an increasingly sophisticated set of mental representations.

Thus a child can acquire CAT via the accumulation of perceptual and causal information of a certain natural kind. This view is thus not committed to radical nativism or to some kind of nonpsychological triggering, since concepts are used throughout the process. Yet the process is more like *pattern* learning or recognition than like hypothesis testing. Ironically, as Laurence and Margolis point out, something like Fodor's own theory of content can actually be used to handle the problem of concept acquisition. This is a very important and frequently overlooked point. A suitably modified causal theory of content can also explain just how the concepts, as mental representations or vehicles, are acquired in the first place. This, in turn, alleviates the need for radical nativism.

This line of thought is closely related to what Carey (2009) calls "Quinian bootstrapping." Carey also provides us with reason for optimism in

explaining our ability to acquire genuinely new atomic mental representations via a psychological process and sustaining mechanism. Preexisting concepts play a causal role in setting up sustaining mechanisms. Even if some of the details of this approach are incorrect, we still have good theoretical reason to suppose that a solution to Fodor's puzzle is within reach.

What role does language play in the acquisition of perceptual concepts? In the case of humans, we should acknowledge that language can aid the process of concept acquisition. Using linguistic labels helps to establish a difference between concepts of distinct object kinds in a complex individuation task. In short, words can facilitate categorization. Language is not necessary for all concept acquisition or possession but can make it easier to acquire some concepts. It is even likely that language is *necessary* for having *some* sophisticated concepts and thoughts, such as my current belief that the speed of light is 186,000 miles per second. Some have argued that language plays an important role in concept acquisition by words serving as "essence placeholders" for objects (Xu 2002). Since Plato's time, it has seemed plausible to many that there must at least be some important underlying similarity among a group of objects when they are called by the same name.[11]

Nonetheless I hesitate to embrace a full-blown psychological essentialism as a criterion for concept possession for two main reasons. (1) Recall that in CONPOSS I used the term "central feature" of a category as opposed to "essential property." It seems to me that an infant can still have *some degree* of the concept TIGER by virtue of observable properties only, such as the way that tigers typically look and move. It may well be that infants and young children do *increasingly* suppose that "hidden" or "nonobvious" properties are taken as a category's primary property. But I do not wish to treat such a hidden property as essential for an infant to possess TIGER. (2) Psychological essentialists also tend to embrace the theory-theory of concepts across the board. It seems to me, however, that we need not do so.

Much the same process described in the previous paragraphs occurs with respect to *artifacts*, such as DOORKNOB or CAR. We begin with innate concepts such as SHAPE, SPACE, and THING. As Cowie puts it for doorknobs, we learn to "respond selectively to [adult] waist-high, movable protuberances attached to doors" (1999, 136). Over time, we become more and more able to recognize and reidentify cars and doorknobs. The important difference here is that *function* and *intentions* are more central to artifacts than to natural objects. One might even suppose that the original intended function of an artifact *is* its essence. Learning about some artifacts might also require

additional attention and motivation, as one is often driven by the development of a practical skill.

But even here, innate concepts such as CAUSE (and EFFECT), MOTION, and SELF are at work from the beginning. A child understands that "if I move this thing (i.e., doorknob) this way, then this other thing (i.e., the door) opens." Although some who sympathize with the theory-theory of concepts have held that it does not apply to artifacts, others argue that the difference between natural kinds and artifacts is not that great. Keleman and Carey (2007) hold that infants and young children develop an understanding of artifacts as objects that are designed with certain initial intentions. Bloom (2000) argues that psychological essentialism holds for artifacts as well as natural kinds.

One interesting result is that nine-month-olds are able to distinguish between little models of birds and airplanes, all of which have outstretched wings (Pauen 2002). It would thus seem that they are at least distinguishing between models of animate and inanimate objects despite the similarity of appearance and lack of any motion. Infants are thus sensitive to the difference between (biological) animals and (inanimate and artifactual) vehicles. On this and other supporting bases, Mandler concludes that "in the second half of the first year a number of global conceptual distinctions are made that differentiate animals, vehicles, furniture, utensils, and plants" (2007, 199). Infants are soon thereafter able to make conceptual distinctions *within* VEHICLES and ARTIFACTS. It is important to note that there may be significant differences between cultures as to which artifacts and animals infants typically encounter. But infants are highly sensitive to motion and thus understand ANIMAL as an object having its own internal source of motion. ANIMATE applies to animals but not to artifacts.

With regard to *property* concepts, matters are a bit different and often more complex. It is again crucial to note that infants, without even realizing it, are distinguishing objects from one another, and it would seem that they must do so by virtue of *some* properties, likely beginning with shape and location. So an infant will acquire some property concepts very early in infancy, such as TEXTURE and specific shape concepts such as SQUARE or ROUND, whereas color concepts seem to develop a bit later. But even if we accept the object-first hypothesis, the infant also begins to acquire quite a number of property concepts early in cognitive development without much conscious effort or attention.

So let us describe the process more generally. Suppose an object O has properties F, G, H, and I. It may be that another object, A, has F and G, but not H or I. Object B might have only F and I. Now these properties might

be specific shapes, textures, or colors. Infants will thus interact often with them via exposure to a number of different objects. Through implicit learning and memory, a child will become increasingly perceptually sensitive to a property, such as roundness. This will in turn allow the child to recognize subsequent round objects, as well as to differentiate round objects from nonround objects. So a child will then be able to apply concepts of F, G, H, and I to objects A, B, and O.

Children can quickly build up a stock of concepts via implicit learning and then eventually with explicit learning and language learning. Since concepts constitute thoughts, children will then be able to form numerous additional thoughts via an inferential process. Although a high level of inferential ability need not be required for having any concepts, it is certainly one important way to acquire additional concepts and knowledge. And solid evidence seems to indicate that by nine months, infants are capable of making inductive generalizations and abstract inferences.[12]

7.3.3 More on Concept Acquisition: Implicit Learning and TILT

To clarify the relationship between consciousness and concept acquisition, more needs to be said about exactly how we can acquire concepts in the absence of explicit conscious attention. Notice that the above account of concept acquisition does not entail that the subject consciously grasp the concepts in question. As we saw in chapter 2, mental content can be acquired at the level of unconscious states. First-order conscious states also require unconscious HOTs or METs.

Thus, given the previous discussion, I suggest that a key notion in explaining how an organism can acquire a concept without either conscious attention directed at objects or brute triggering is *implicit learning*. What makes this process of concept acquisition still a bit mysterious is that we seem able, from a very early age, to acquire concepts without paying much explicit conscious attention to the referents of those concepts. The basic idea behind implicit learning is that it involves learning that guides the subsequent behavior of an organism O, but where O is not aware that O had learned anything and is not aware of what O had learned (Kihlstrom, Dorfman, and Park 2007).[13] Implicit learning is basically *unconscious* learning and is often contrasted with *explicit* learning, which is more like Fodor's sophisticated notion of conscious hypothesis testing. We might therefore refer to this theory of infant concept acquisition as The Implicit Learning Theory, or TILT.

One oft-cited study of implicit learning comes from early work on artificial grammars (Reber 1967). Subjects were asked to memorize lists of letter

strings, such as "XXRTRXV" and "QQWMWQP," that are generated by rules unknown to the participants. Participants are then told that the strings they memorized followed certain rules, and are asked to classify new strings as grammatical (i.e., following the rules) or not. The subjects typically perform much better than chance at this classification task, indicating that some knowledge has been acquired, despite the subjects' being unable to verbally describe or consciously access the rules.

In the case of implicit *concept* learning, subjects are able "to identify instances of novel concepts . . . without being able to describe the defining or characteristic features of the concepts themselves" (Kihlstrom, Dorfman, and Park 2007, 535). A subject can learn something new but not know what they know. To be clear, the claim is not that the *subject* is unconscious but that there is no state consciousness of the process of concept acquisition. Thus implicit learning is typically construed as *unconscious* in at least this sense, but it is also a psychological process.

For example, Knowlton and Squire (1993) seem to have demonstrated that memory-impaired patients had an intact ability to learn a novel category. In this case, the category had to do with a prototype dot pattern whereby subjects implicitly learned to judge whether or not subsequent patterns are members of that underlying prototype. Perhaps most relevant to our purposes is a study by Reed et al. (1999), where the items that were used were artificial animals. As opposed to abstract dot patterns, these items are not difficult to describe in terms of familiar notions, such as body, neck length, and shape. Similar results to the dot pattern study were found, that is, intact categorization with impaired recognition. Although there are interesting and legitimate concerns regarding some of this literature, it seems that the overall case for implicit "category" learning is strong (E. Smith 2008).

Once again, we do also sometimes explicitly, and thus consciously, learn concepts. This is especially true later in life or in more formal settings, such as in school or a home teaching environment. And when one is capable of introspection, especially in adulthood, more sophisticated and complex learning occurs, such as learning various concepts in physics, law, philosophy, or accounting. But the main difficulty we are concerned with, Fodor's puzzle of concept acquisition, centers more on how we acquire the plethora of *empirical* or *perceptual* lexical concepts in infancy and childhood. It is not as if children in elementary school or even at home are normally *explicitly* taught what a table, a tiger, or a doorknob is. This is where implicit learning and TILT can help to explain the data and the rapid acquisition of concepts in infancy.

It is worth mentioning that implicit learning is often discussed in connection with (or by analogy to) implicit *memory*, as opposed to the more explicit and conscious *episodic* memory (Kihlstrom, Dorfman, and Park 2007). Implicit memory is basically unconscious memory, that is, memory acquired without awareness of the memory. *Repetition priming* would be a classic example of implicit memory. Repetition priming involves the facilitation in the processing of a stimulus as a result of a recent brief encounter with it. And memories are, after all, also mental representations, so the relationship between implicit memory and implicit learning is a strong one.

Part of the value of this comparison, as I alluded to earlier, is that there are numerous rather odd cases of implicit learning in amnesiacs who do not even consciously remember the learning sessions at all but have preserved knowledge from implicit learning. For example, there are instances where an amnesiac performs better over time on a maze learning task despite having no conscious recollection of previous test trials. This would be an example of *procedural* memory, whereby one acquires a skill ("knowing how") as opposed to some kind of *declarative* knowledge (or "knowing that"). Moreover, there seems to be evidence of implicit learning in Alzheimer's patients, despite the memory impairments (Bozoki, Grossman, and Smith 2006). There is also evidence of different brain areas responsible for implicit and explicit memory or knowledge (Reber et al. 2003).

We need to be careful, however. There is some ambiguity in the notion of implicit memory. It is most often defined in terms of the *product* or knowledge gained, rather than the learning *process* itself. Of course, in the case of severe amnesiacs, there is no explicit memory of either the process or the result. But even when being formally tested, subjects are not always aware that they have learned something. So although subjects will normally later remember *the period of testing* itself, they still do not realize that they were learning something specific during the process. This indirectly shows the value and plausibility of implicit learning in early concept acquisition. As we have seen, infants and young children seem to be able to learn concepts at an incredible speed, but there is little evidence of conscious attention directed at the objects or explicit memory at a later time. Indeed, this is precisely what leads some to embrace a form of the poverty-of-stimulus argument for radical nativism.

Furthermore, we have increasing evidence for what is now called *implicit working memory* (Hassin et al. 2009), as opposed to the view that working (or short-term) memory is always conscious and explicit (Baars 1997). The idea here is that working memory can operate unintentionally and outside conscious awareness. We are aware of only a subset of the information

actively maintained in working memory. For example, the process of *pattern extraction* seems to occur outside conscious awareness. It is precisely this process that enables a subject to acquire a concept over time based on exposure to similar stimuli.

Mandler (2004, 2008) posits what she calls the perceptual meaning analyzer (PMA) mechanism, which provides a redescription of patterns in perceptual data. But this is "an *attentive* process that extracts spatial information from perceptual displays and . . . recodes it into a skeletal . . . form" (2008, 212; italics mine). We might think of it as a data summarization device. The obvious problem, however, is that even if there were such a mechanism, it cannot explain *implicit* learning in infants if it requires a consciously attentive process. Early infant concept formation is the main puzzle, and there is little reason to think that young infants pay as much attention to objects or properties under normal circumstances as they do in experimental contexts. Indeed, Mandler basically acknowledges that there is no real compelling evidence for the existence of PMA and that we do not routinely attend to much perceptual information. "A great deal of perceptual information is taken in parallel and most of it is processed outside of awareness" (2008, 212).

On the other hand, Mandler explains that infants also "form *perceptual schemas* of objects, so that dogs and chairs and so forth become familiar objects to them. However, perceptual schema formation is *implicit learning* of similarities that requires neither attention nor awareness" (2008, 211–212; italics mine). Thus this process is automatic and typically occurs *without* attentive control of the perceiver, which is more in line with implicit learning. However, she describes the content as "analog," suggesting that it is nonconceptual, which would be at odds with the notion that genuine *concepts* can be acquired via an implicit learning mechanism.

Another contentious issue stems from Mandler's distinction between "perceptual" and "conceptual" categorization. Using her own somewhat peculiar terminology, perceptual categorization refers to objects and properties that are observable features (such as "has four legs" or "is red"), whereas conceptual categorization refers to more abstract and nonobservable properties of objects (such as "continuing to exist when unperceived" or "being a self-propelled agent that causally interacts with other things"). Conceptual categorization includes background knowledge and information about, say, the ontology, causation, and function of objects. Mandler argues that there is a clear distinction between these two types of categorization and that infants have two distinct systems that operate in parallel from very early in infancy.

This is a controversial and, I think, confusing distinction, and many researchers claim that we cannot draw such a clear line between perceptual and conceptual categorization or that perceptual categorization also involves what Mandler calls conceptual categorization. Based on other experiments, Rakison (2006, 2007), for example, argues that many of the same objects or entities that one might categorize as "self-propelled" or "animate" (dogs, animals, birds) also have certain characteristic observable properties (legs, wings). This perhaps also calls into question the notion that infants use hidden or nonperceptual concepts before or parallel to perception-based properties. Infants might successfully categorize animals as different from vehicles because they attend to specific (perceptual) object features and not because they possess innate or superordinate concepts like ANIMATE. Infants may learn over time that there are statistical regularities between perceptual features of objects (legs, wings, wheels, eyes) and whether they tend to move with or without external cause. If infants really are successfully categorizing, say, animals and vehicles on the basis of perceptual features or parts (such as THING WITH LEGS), then the *order* of concept acquisition may not be what Mandler and others have argued.[14]

In any case, I have argued here that so-called percepts and concepts are interrelated in a conscious state. Indeed, what are sometimes called "percepts" are concepts in my view; that is, they are concepts of observable properties and objects. This is a view that clearly has affinity to conceptualism and the WIV. Conscious perceptual experience involves two intertwined levels of concepts, one on the first order and one on the higher order.

Overall, then, my TILT account of early concept acquisition relies on view that there are some initial general core (innate) concepts. Additional more specific concepts are acquired via an unconscious (or implicit) psychological process with the help of implicit learning, a sustaining mechanism, and a causal-development theory similar to Rupert's (1999) account.

7.4 Conceptualism and Concept Acquisition

Recall the following two difficult questions: How can we acquire concepts if conscious experience presupposes having concepts (or conceptualism is true)? How can one acquire concepts if conscious experience itself involves a HOT (with its constitutive concepts)? In this and the next section, I argue that concept acquisition and infant consciousness are consistent with both conceptualism and HOT theory. The keys lie in the already developed notions of "perceiving-as" and the importance of implicit learning in TILT.

In some ways, then, this section is a natural extension and continuation of section 7.3, especially given my overall goal of showing how one can acquire concepts without presupposing prior conscious understanding of those concepts.

7.4.1 A First Sketch of a Solution

We begin in infancy by experiencing objects (or properties) in a certain way, or *as* certain things, employing a select few core concepts. This allows conscious experience to get started in a way that is consistent with conceptualism. As we acquire more and more concepts via implicit learning, we are able to experience additional objects (or properties) *as* falling under these concepts. One's stock of (sensory or perceptual) concepts builds up very quickly. As we saw in the previous chapter, however, we can still *see* an object O with property F without first having the concepts for O or F. But I urged that *at that initial time*, we need not be experiencing O *as an O* or F *as O having an F*. The infant might, for example, initially see the trees only *as* THOSE MANY LARGE DARK CIRCULAR OBJECTS. These concepts involve many innate primitives, such as NUMBER, OBJECT, and SHAPE. Through repeated exposure to the objects or properties in question, we acquire more and more concepts, which, in turn, we can apply to future experiences. Thus there is a three-step process: (1) we see an object O (or a property F); (2) we acquire the concept of O (or of F); and (3) we experience O *as* an O (or *as* having F).

The most difficult part to explain is step 2. TILT is designed to help at precisely this stage. Implicit learning and memory are the keys to understanding how infants move from step 1 to step 3 without requiring that the subject perceives an object O *as an O* during steps 1 or 2. As we will see, this will help to disarm two interrelated arguments against conceptualism raised by Adina Roskies (2008, 2010).

7.4.2 Roskies's First Argument

Roskies does a nice job of setting out some of the key issues discussed in this and the previous chapter. But she argues that conceptualism "entails concept nativism. That concepts are innate, not learned, is conceptualism's price, and it is too high a price to pay" (2008, 634). Roskies normally has in mind radical (or "widespread") concept nativism, and she argues that the conceptualist cannot account for concept learning. Thus "nonconceptual content of experience must be invoked to account for concept learning" (636). Needless to say, I strongly disagree with Roskies and take much of the preceding discussion to be a refutation of her overall argument. But let us look more closely.

Roskies offers what she takes to be a *reductio* of conceptualism; that is, assuming conceptualism is true, we are led to the absurd conclusion that *all* perceptual concepts cannot be learned. Using RED as her prime example, Roskies argues that *either* having red experiences must already involve RED (in which case there is no explanation for how RED was acquired) *or* having red experiences does not involve RED (in which case RED is, implausibly, built up from other concepts the thinker possesses). Thus RED is not learned, and the same is true for all other lexical perceptual concepts.

Given the discussion of this chapter, we can see that two lines of reply are open to the conceptualist. First, we might suppose that RED does *not* occur in the content of experiences an infant has when it *first* sees red things (much as we have seen by analogy with PYRAMID, TREE, and so on), but it does not follow from this that RED is itself "built up out of other concepts." Rather, an unconscious psychological process (in this case, very quickly, using innate and already present concepts like SPACE and COLOR) results in the possession of RED, after which a similar infant conscious experience will have RED as part of its content. Thus we do have an explanation for how lexical concepts can be acquired without supposing that the same concepts are already present in conscious experience. Remember that this is very early in infancy.

Second, as we have also seen, Fodor's two ways of acquiring a concept are not exhaustive, as Roskies seems to recognize later in the paper. Yet Roskies, for most of her paper, still follows him in supposing that there are only two ways to acquire a concept, namely, either by some kind of fairly sophisticated and person-level concept learning involving effort and attention (albeit not necessarily by using hypothesis generation and testing) or via a brute causal nonpsychological process that merely triggers an innate concept (Roskies 2008, 642–643). But a third possibility is that many concepts are acquired by a cognitive, but unconscious, process that does not treat RED as compositionally built up out of other concepts. We have already seen how this might go in the previous section with the help of implicit learning and TILT. In this case, we have an acquired property concept. But, again, lexical concept acquisition is achieved via an unconscious psychological process. The acquisition of a concept C can still be a "temporally extended process" and a "cognitive achievement," to use Roskies' terms, but still not involve a personal-level experience of the relevant object or properties *as* being C.

In addition, her paper shows some ambiguity with respect to terms such as "learning," "acquisition," and "attainment." As we have seen, it is crucial to be clear about these notions. Although Roskies tells us that Fodor's

notion of "learning" is too restrictive, she also follows him in supposing that concept learning must be a personal-level phenomenon. It is fine to restrict "concept *learning*" in this definitional way, but then we must also acknowledge the possibility of unconscious concept *acquisition* based on TILT.

Conceptualism may certainly require *some* innate concepts, such as those in the core concept nativism developed earlier. But I take it that Roskies is most anxious to saddle the conceptualist with a much stronger form of nativism. Indeed, she says, "I do not wish to argue that any degree of nativism is sufficiently troubling to warrant the abandonment of conceptualism; some concepts may well be innate" (646). But I contend that Roskies is mistaken when she later insists that "core concept nativism seems to rely upon the admission of nonconceptual content . . . [and thus] limited nativism . . . is not an option available to the conceptualist" (646). Indeed, a central theme of this and the previous chapter has been to show otherwise. Of course, the importance of Roskies' paper is that concept acquisition should be treated as a central and difficult challenge for conceptualism. I certainly agree with this.

To her credit, Roskies does mention the possibility of implicit learning as an objection to her overall argument (2008, 652–654). But she dismisses this option much too quickly and does not close off the option of unconscious concept acquisition via something like implicit learning.

First, Roskies argues that although implicit learning is a genuine phenomenon, concept learning must still be a personal-level phenomenon. But, as we have already seen, there is a middle ground between personal-level concept learning, on the one extreme, and noncognitive brute causal processes of concept acquisition, such as triggering. The main point, again, is that just because a subject is unaware of a causal process that results in concept acquisition, it does not follow that the process in question is nonpsychological (or noncognitive). Contrary to what Roskies claims, the process that eventually results in a subject being able to deploy newly acquired concepts need not be so mysterious. She is right, however, in saying that the conceptualist still "owes us a plausible story for how the sub-personal representation is made available to the thinker" (654).

Second, Roskies thinks that invoking implicit learning in this context relies on a misunderstanding of what it is: "Implicit learning involves the learning of an association between stimuli in such a way that the subject is unaware *that she is learning that association*. It is not a phenomenon in which the subject is unaware of the stimuli themselves" (654). Thus, in the end, she doesn't think that the data from implicit learning threaten her argument. But the key question is: is it true that when one implicitly learns

some concept C, one is consciously aware of the object or property in question *as a C*? This is where I think the answer is no, as we will see even more clearly in the next subsection. But we have already seen that conceptualism need only require that we can, for example, see an object O with property F without first having the concepts for O or F. But then *at that time* we would not experience O *as an O* or F *as O having an F.* The conceptualist can rightly insist that during the process of implicit learning, one is not applying the eventually acquired concept to the stimulus in question.

Finally, it is not at all clear that implicit learning *always* merely involves learning an "association." Kihlstrom, Dorfman, and Park (2007, 535), for example, list "association" or "covariation detection" as only one of four varieties of implicit learning. Sequence learning is also listed, but so is "concepts" where "subjects learn to identify instances of novel concepts . . . without being able to describe the defining or characteristic features of the concepts themselves" (535). Implicit learning "goes beyond the formation of simple associations . . . and involves the acquisition of knowledge of some complexity, at some level of abstraction" (535). Indeed, it is difficult to see why Roskies would acknowledge that implicit learning occurs and that one is able to acquire an *association* between a mental representation and a perceptually represented object, but then urge that implicit learning somehow cannot result in the concept acquired in the first place. It seems to me that the former is more complex than the latter.

Roskies also considers the possibility that her opponent might adhere to what A. D. Smith (2002) calls a "low" (or minimal) theory of concepts, according to which concept possession is merely having a certain ability to act differentially with regard to a set of entities, such as merely by discrimination. She rightly points out that if one holds too "low" a theory of concept perception, then conceptualism becomes trivially true, and what she calls "nonconceptual" becomes "conceptual." But then she goes on to point out that the debate in question involves those who typically hold the more sophisticated "high" theory, according to which mental representations are invoked to link concepts to a range of properties or objects in the world, to think about those things or properties, and perhaps even to the capacity for language use or abstract thought. Roskies concedes that her argument "applies only to high theories of concepts" (649) but notes that at least one conceptualist holds a "low" view (Noë 2004).

But first, while it might be right to say that *most* conceptualists, such as McDowell, typically have a "high" theory of concepts, it clearly doesn't follow that all do or that a high view of concepts is essential to conceptualism. Second, although the criteria of concept possession as expressed in

CONPOSS are clearly not "very high," I have also suggested that the criteria are also not too low, either. We might label CONPOSS a "medium theory" that is strong enough to avoid the charge of triviality but weak enough to avoid Roskies' argument. My view of concept possession is lower than others; for example, it does not require language use. It is also lower than others because it importantly allows for degrees of concept possession. On the other hand, it is higher than a trivial account that would allow for mere discriminatory ability to be sufficient for concept possession. Once again, there is a middle ground ignored by Roskies.[15]

Roskies is also rightly concerned about infants and animals in connection with conceptualism. Here I disagree with Brewer and McDowell, as we will see later in this chapter and the next. In essence, they are giving up on solving the Consciousness Paradox. Brewer and McDowell deny that infant cognition is conceptual, but they also seem to hold that concepts are learned without offering a worked-out theory of concept learning. "Conceptualists tend to think that the mental lives of humans and animals differ in kind, whereas nonconceptualists see human and animal mental lives as continuous" (Roskies 2008, 650). Thus, with respect to the relationship between human and animal minds, I may not only differ from "typical conceptualists" but am also concerned to defend a version of HOT theory. Indeed, a central theme of this book is to show that conceptualism, HOT theory, and animal and infant consciousness are jointly consistent.

7.4.3 Roskies's Second Argument

The reply in the previous section can also, in part, be used against yet another anticonceptualist argument from Roskies (2010). She argues that demonstrative concepts cannot be used as a panacea to reply to nonconceptualist arguments. As a matter of fact, she argues that demonstrative concepts actually undermine their purpose because they presuppose nonconceptual content. Here is her formal argument:

(1) Forming a demonstrative concept requires a demonstration.
(2) The relevant demonstration in conceptual demonstrative formation is the endogenous (voluntary, intentional) focusing of attention.
(3) Intentional focusing of attention involves representational content of experience.
(4) To be a response to the learning argument, that representational content cannot always already be conceptual.
(5) *Thus*, forming a novel demonstrative concept appropriate to account for novel concept learning must involve focusing attention on contentful aspects of experience that are nonconceptual. (Roskies 2010, 123)

A few preliminary remarks first: (a) Roskies is correct that premise (1) is unproblematic. I will not challenge it. (b) Roskies is also correct that demonstrative concepts cannot so easily handle the problems for which they are invoked. As I mentioned in the previous chapter, I am not as enamored with the demonstrative concept strategy as McDowell or Brewer. There remains the further problem that I called "the priority argument," which is central to an account of concept acquisition (and so for demonstrative concepts as well). Nonetheless much of this chapter (especially the previous subsection) has been dedicated to showing that premise (4) is false. So we have already responded to Roskies' specific charge that learning requires nonconceptual content (in conscious experience) in what she refers to as the "learning argument." Once again, the conceptualist is not forced to embrace radical nativism. (c) Roskies reiterates her contention that conceptualists use a "high" theory of concepts. But we have seen how a conceptualist can embrace a weaker, though not minimal, theory of concepts in contrast to McDowell and others.

Now, Roskies spends most of her time treating premise (3) as the most controversial premise. This is understandable, but it seems to me that premise (2) is more problematic. Once we see why, we will also see that other replies are available to the conceptualist. Roskies tells us that a cognitive analogue to pointing (i.e., demonstrative reference) is the focus of attention. But she clearly means *conscious* (and so "personal-level") attention, which involves the "voluntary" and "intentional" focusing of attention. However, this causes a real problem for premise (2). Just as we have shown how a conceptualist can use implicit learning in response to her learning argument, so we can now argue that the demonstrative reference in question (and thus demonstrative concept formation) can be secured without conscious attention and thus without *conscious* nonconceptual content. That is, while it is true that *one way* to secure demonstrative reference is via conscious attention, there are other relevant ways as well. Let me explain.

We have already conceded that there may well be *subpersonal* nonconceptual content (sec. 6.1.3) that could never come to consciousness, such as is the case in early visual processing. Recall the views of Pylyshyn (2007), Raftopoulos and Müller (2006), and Fodor (2007), all of whom support the notion that there is subpersonal informationally encapsulated nonconceptual content. If a case can be made that the "demonstration" in demonstrative concept formation need not involve conscious attention, then this would falsify premise (2) and bolster the claim that fixing the reference of a demonstration can be achieved unconsciously. Although Roskies acknowledges that "attention can be automatically drawn to salient stimuli"

(2010, 124), she insists that conscious or voluntary attention is required for demonstrative concept formation. She relies heavily on Campbell's view that attention must operate at the level of experience; that is, something in conscious experience must make possible the formation of a demonstrative concept: "Appeal to the agent's demonstrative intentions requires us to appeal to the agent's conscious attention to objects" (Campbell 2002, 14).

But first we have already argued that embracing "deep" subpersonal nonconceptual content is consistent with conceptualism *in conscious experience*. Second, Roskies seems to rely on the assumption that no causal theory of mental content could satisfactorily fix reference without invoking personal-level or conscious attention. She approvingly cites Campbell's view that "to ground reference[,] perception must involve an experiential component" (Roskies 2010, 126), and says that "we must have conscious access to content which is sufficient to distinguish one of the objects from the other" (130). At the least, this is a substantial assumption that is far too premature to make (see also chapter 2). So we should not rule out demonstrative reference without conscious attention.

Third, to use a specific case, one might point to Pylyshyn's work in this regard. Recall that he uses subpersonal nonconceptual content to account for perceptual reference. He discusses "pointers" directed at external objects that function like demonstratives that he calls FINSTs. Such demonstrative pointers early in the visual system allow an organism to parse the visual world and so segregate things in space and time. Thus, if Pylyshyn is right, we do not need conscious attention to track and fix specific reference to objects. The sort of direct connection that he has in mind is also very much analogous to linguistic demonstratives like "this" or "that" and uniquely picks out particular individuals. There is a visual mechanism that automatically selects and indexes a number of visual objects within one's visual field.

Fourth, and perhaps most important, Campbell's (and thus Roskies') view has been seriously challenged in more recent literature and, to my mind, in some extremely convincing ways. For example, Matthen (2006) shows in greater detail just how reference to objects can be established through visual processing in the dorsal visual stream. We have already seen how dorsal-stream representations can be genuinely intentional (and have concepts) and yet unconscious. Reference can be determined unconsciously. So Campbell does not give us sufficient grounds for thinking that consciousness is *always* involved in visually attending to and picking out unique objects.[16]

More recently, Raftopoulos (2009b, 350–359) further criticizes Campbell and argues that identifying reference to objects does not require conscious

visual attention but only requires the *preattentive* perceptual mechanisms of object segregation. For example, he offers significant empirical evidence for the view that object segregation takes place at many different levels of visual processing, including an early and purely bottom-up stage before concepts are applied. There are, to be sure, some additional thorny issues involved, such as whether genuine sortal concepts and spatial attention (or location) are needed to establish demonstrative reference. But my point here is mainly that premise (2) is not adequately supported, and Roskies does not address a number of important challenges to it.[17]

To her credit, however, Roskies does briefly anticipate something like the foregoing reply toward the end of her article, but then dismisses it much too quickly and offers weak replies on behalf of the conceptualist (2010, 128–130). She insists that any such reply involves some "obscure process" or "fancy footwork." I hope I have shown throughout this chapter that this line of reply is much more plausible than Roskies supposes. The bottom line, then, is that we need not hold that conscious attention is the only way to form demonstrative concepts. Thus demonstrative concepts, to the extent that we do have them, need not presuppose nonconceptual content *in conscious experience via conscious attention*. Instead they could be formed or acquired via unconscious attention with unconscious reference fixing.

Now, when such content is genuinely conceptual via implicit learning, the subject can then become aware of that content, which, according to the HOT theory, is precisely when state consciousness is present. Indeed, implicit concept learning can be achieved without conscious attention to, or focal awareness on, objects or referents. Thus, even if forming novel demonstrative concepts requires nonconceptual content at some deep level of visual processing, this kind of nonconceptual content is not a threat to conceptualism. For example, forming the demonstrative concept THAT SHADE or THAT SHAPE as deployed in conscious experience need not involve them as having been already present in conscious experience or conscious attention. And the reason that deep subperceptual content is nonconceptual has to do with our criteria in CONPOSS. At this very early stage of visual processing, there may well be a crude ability to discriminate between (or "segregate") objects via certain properties. But there is no conceptual *recognition* of the objects or properties, not to mention that no genuine intentional states are formed at this stage. Moreover, the kind of subpersonal nonconceptual content in question can also involve highly sophisticated notions, such as the ones we find in Marr's theory. Such deep nonconceptual content is closed off from becoming part of a HOT, which would enable a subject to have conscious states with that content.

One might reply that the previous line of argument undercuts some of my initial rationale for accepting core innate concepts, such as OBJECT, NUMBER, or MOTION, because of the reliance on deep subpersonal nonconceptual content. I disagree. First, the foregoing reply using dorsal-stream conceptual representations would still hold even if there were a problem otherwise. If we accept the two-systems hypothesis, it would seem that conceptual demonstrative reference can also be achieved via the dorsal (unconscious motor) pathway. This alone suggests that conscious attention is not required to fix reference of the external objects in question.

Second, the rationale and evidence for accepting the core group of innate concepts earlier in this chapter had to do with the criteria in CONPOSS and the lack of evidence that those concepts are ever acquired. That rationale is unaffected by any plausible *additional* evidence for nonconceptual content at a deeper subpersonal level, such as Marr's "zero-crossings" or Pylyshyn's FINSTs. We have little reason to rule out the formation of genuinely (unconscious) mental states, such as unconscious perceptions, whose content involves demonstrative concepts. Indeed, we even noted that there is good reason to include demonstrative concepts, such as THIS and THAT, in our list of core concepts. As we saw, core innate concepts serve to discriminate between (or segregate) objects, identify objects and properties, and form the basis of early primitive thoughts. These are all criteria in CONPOSS.

Finally, we saw in chapter 4 that many psychologists do allow for unconscious attention. At the least, the matter of just how attention and consciousness are related is somewhat more complicated than Roskies recognizes. It is surely open to a conceptualist to allow for unconscious demonstrative concept formation. At any rate, I conclude here that Roskies' arguments fail to refute conceptualism.

7.5 HOT Theory and Infant Consciousness

It is widely held that infants are conscious, at least with respect to simple emotions and feelings, and that there are levels of consciousness (Trevarthen and Reddy 2007; Zelazo, Gao, and Todd 2007). Thus, as a stand-alone claim, the Infants Thesis is fairly uncontroversial. Alison Gopnik (2009, chap. 4) goes further arguing that babies might even be *more* conscious than adults in the sense of having a heightened, though less-focused, outer-directed awareness. She bases her view on both neurophysiological and developmental findings, such as the fact that infant brains have fewer inhibitory neurotransmitters than the brains of older children and adults. She uses the clever metaphor of a "lantern," instead of the more focused

attentional "spotlight" of consciousness for adults, whereby infants more vividly experience everything at once rather than experiencing a single aspect of the world. We might say that, for Gopnik, infant experience is perceptually *richer* than adult experience in the sense that more is experienced (though presumably with fewer fine-grained concepts). Gopnik argues that since there is so much novelty in a baby's world, we might compare such consciousness to the heightened outer awareness of an adult on a trip to an unfamiliar country. As we have seen with the habituation technique, babies also reliably look longer at unexpected or novel events. This technique actually gets harder as babies grow older because their attention becomes more controlled by an internal and voluntary agenda.

However, given the demands of HOT theory, some might question whether infants can be conscious if HOT theory is true. Indeed, infant and animal consciousness is often construed as a problem for HOT theory. Animal consciousness will have to wait until the next chapter. Recall that HOT theory has it that a thought of the form "I am in M now" must accompany all conscious states. Let us first tackle the "I-concept" in HOTs and then the mental concept ("M") in the following subsection. I think that HOT theory is perfectly compatible with infant consciousness (and perhaps even late fetal consciousness).

7.5.1 Infant Consciousness and I-Thoughts

We have already seen some evidence for a primitive and innate I-concept (SELF) present at birth. For example, infants are able to distinguish their own bodies from outer objects (Rochat 2001). And it is crucial to recognize that there are varying degrees of "self" or "I" concepts (Rochat 2003; Morin 2006). More generally, a number of authors have argued that there is a primitive *bodily self-awareness* that involves some form of self-concept (Gallagher 2005; Legrand 2007a, 2007b). Some researchers stress the importance of proprioception as a primitive form of bodily self-consciousness in infants, partly based on the well-known innate rooting response where infants tend to orient their head toward touch stimulation (Rochat 2003; Legrand 2007b). Such a rooting response may not *by itself* suffice for concept possession or self-consciousness, but other evidence is forthcoming. A number of essays in Zahavi, Grünbaum, and Parnas 2004 also argue for the existence of a primitive, implicit, and bodily form of self-consciousness. For example, very young infants are able to discriminate and identify their own leg movements displayed in a mirror from those of another infant (Jeannerod 2004).

It is also worth noting here the related point that infants may also possess a self-concept based on deployment of an innate concept CAUSE as the *source* of various bodily movements. Jeannerod (2004), for example, argues that sensory-motor mechanisms allow us to recognize our bodies and our actions *as our own*. I develop this idea further in the next subsection because it is also relevant to the notion that AGENT, or some mental concepts, should be added to the list of innate concepts. Based in part on the pioneering experiments in the 1960s by Nielsen (1963), Jeannerod suggests that infants have the ability to recognize themselves and others as agents of behavior. That is, infants can at least distinguish between self-generated actions and actions produced by others. A self-generated action brings with it both a sense of ownership and a sense of authorship. Moreover, infants react differently to people than to inanimate objects.

In any case, with respect to HOT theory, it is crucial to remember that having an implicit or primitive concept of self need not involve what we have called *introspection* or *reflection*. According to HOT theory, an organism can have conscious states as long as it has *unconscious* HOTs. Any evidence against the possibility of infant *introspection* does not entail a lack of state (or creature) consciousness.[18]

I have also argued elsewhere (initially in Gennaro 1993) that there are degrees of self-concepts. We should distinguished the following four I-concepts, moving from the least to most sophisticated:

Level 1: I *qua* this thing (or "body"), as opposed to other physical things.
Level 2: I *qua* experiencer of mental states.
Level 3: I *qua* enduring thinking thing.
Level 4: I *qua* thinker among other thinkers.

Some of these levels of I-concept are similar to what Rochat (2003) and Morin (2006) seem to have in mind. And all the authors mentioned in this subsection agree that infants are at least capable of having a level-1 self-concept.[19]

I suggest that we also have additional evidence for infant self-concepts based on recent work on episodic memory in infants. Episodic memory is an explicit, conscious, and autobiographical form of memory, distinguished from implicit and procedural memory. But episodic memory is not merely a memory of a past experience. There is also an implicit self-concept involved. When I episodically remember going to the Who concert in 1990 or watching a movie last night, I experience (or perhaps *re*experience) it as part of *my* past (Tulving 1983).

A number of empirical results strongly suggest that infants not only have "working" or "short-term" memory but even longer-term episodic memory (Rovee-Collier, Hayne, and Colombo 2001; Oakes and Bauer 2007). For example, using a variation of the familiarization procedure, investigators have found that infants can recognize their mother's face and discriminate it from the face of another woman within three to four days after birth (Pascalis et al. 1995; Rovee-Collier, Hayne, and Colombo 2001, 101). Much the same can be said for recognizing their mother's voice and discriminating the smell of their mother's breast milk. In what are called "deferred imitation" tasks, Meltzoff showed that nine-month-olds can imitate an experimenter's unique actions after a twenty-four-hour delay (Meltzoff 1988), and fourteen-month-olds can exhibit deferred imitation after a four-month delay (Meltzoff 1995). Nine-month-olds remember individual actions over delays of as many as five weeks (Carver and Bauer 1999, 2001). Moreover, a case can be made that any alleged analogy between infants and amnesiacs seems to be at odds with the facts (Rovee-Collier, Hayne, and Colombo 2001, chap. 6). Even newborns exhibit retention under test conditions that adult human amnesiacs do not.

Research using a "mobile conjugate reinforcement paradigm" shows that by three months of age, infants exhibit retention of an event that occurred on a single prior occasion. The mobile conjugate reinforcement paradigm basically first involves infants learning to kick to activate a series of mobiles with numbers on each block and then being provided with information that a stained-glass-and-metal wind chime was movable. When infants were shown the stationary wind chime five days later, they kicked vigorously, attempting to move it in the same way as they had the block mobiles. However, infants who had previously viewed the wind chime while it was motionless did *not* attempt to move it during the test.[20]

Some of these findings are somewhat controversial, but if this line of argument is correct, then infants might even be credited with level-3 self-concepts. Infants understand that they are enduring, thinking things because they understand that they have had past experiences. Such understanding requires thoughts, not merely momentary discrimination between, or recognition of, objects. Moreover, infants would thus surely be capable of having conscious states. As we have seen, episodic memory is usually *defined* as a conscious state. And having some kind of self-concept dictated by HOT theory does not decrease the likelihood of infant consciousness as long as the concept can be possessed unconsciously in a HOT. It is certainly true, however, that an infant's sense of the past (and future)

increases during early development. If we apply CONPOSS to the experimental results, it would seem that infants have concepts like SELF, TIME, CAUSE, and AGENT. If so, there is little reason to resist the conclusion that they have some degree of I-concept.

7.5.2 Infant Consciousness and Mental Concepts

Perhaps the more difficult problem for HOT theory arises with respect to *mental* ("M") concepts, such as BELIEF, DESIRE, INTENTION, and PERCEPTION. But just how sophisticated does a mental concept have to be for an infant to possess it?

Recall that we have already alluded to the idea that infants have some concept of an intentional AGENT, which may even be innate. Based in part on Nielsen (1963), Jeannerod (2004) suggests that infants have the ability to recognize themselves and others *as* agents of behavior. That is, infants can at least distinguish between self-generated actions and actions produced by others. Animate objects have the power of self-efficacy, whereas inanimate objects do not. If this is correct, then infants would seem able to *recognize* others as agents capable of intentional action, or at least as authors of some goal-oriented behavior. The idea is that just as infants can distinguish CAUSE (mental state) and EFFECT (bodily movement) within themselves, so they make the same distinction when encountering other people or organisms (see also Carey 2009, chap. 5). As Zahavi (2004b, 44) rightly explains, an infant will recognize when there is both volition and proprioceptive feedback present. In these cases, there is a self-caused or "self-willed" action. However, when neither is present, there is an "other-willed" action, that is, the action of another. And when "the proprioceptive feedback is present, but the experience of volition is absent (as in the case where the mother is moving the hand of an infant), we have an other-willed action of self" (Zahavi 2004b, 44). These considerations lead me to think that infants can thus form thoughts about others and themselves.

Thus the concept AGENT is an important aid in developing knowledge about other minds, as opposed to other objects in general. Further evidence along these lines comes from studies on "threesome intersubjectivity." For example, Fivaz-Depeursinge, Favez, and Frascarolo (2004) argue that infants have a sense of a shared mental world when playing with both parents. In addition to imitation, there is important *three-way* interaction. For example, a young infant who is sharing pleasure or interest with one parent might turn to the other to share her affect with him or her too. This suggests that infants are even capable of level-4 self-concepts and thus of mental concepts as well.

Another line of supporting evidence comes from the literature on "joint attention" (Eilan et al. 2005). By the end of the first year, when an infant and another person are aware of some object or event, an infant can recognize that the other person is also aware of the object or event. This kind of "mutual awareness" manifests itself in gaze following, pointing to objects, and facial recognition. This can be taken as an indication of a developing child's understanding of attention and other minds.

Much of the foregoing argument is bolstered by Johnson (2005), who reviews additional evidence in favor of attributing mental concepts to infants. For example, she discusses evidence that infants attribute perceptions to people. Seven-month-olds looked longer when a moving person collided with another person than when inanimate objects collided in a similar way. And twelve-month-olds looked longer at a person who smiled at one object but then picked up a different object than at a person who smiled at and picked up the same object (Phillips, Wellman, and Spelke 2002). There seems to be an understanding that another person, or agent, is acting on the basis of goals and intentions. Having goals and intentions is also interrelated with grasping the concept DESIRE. If infants recognize when others have goals, then they understand that others WANT to do something, which is a central feature of DESIRE.

Important related results are reported by Behne et al. (2005). Infants as young as nine months reacted with more impatience (e.g., by reaching or looking away) when an adult was unwilling to give them a toy (e.g., teasing the child) than when she was unable to give it (e.g., accidentally dropping it). This indicates that infants understand goal-oriented action at a very early age in that they appear able to grasp the intentions of the agent.[21]

Thus it seems reasonable to suppose that infants are at least capable of some rudimentary ability to attribute mental states to others. It is important, once again, to remember that there are degrees of self-awareness and concept possession. We build up our stock of mental concepts from early and innate concepts, such as CAUSE, EFFECT, SAME, THIS, and TIME, among others. These concepts are basic and coarse-grained at this early stage of development.

Moreover, perhaps concepts such as HUNGER and PAIN are also innate. If not, we can see how they are acquired very early via the application of the innate concepts in the previous paragraph. An infant (or even fetus) can quickly acquire a (primitive) concept of PAIN by applying THIS or THAT to certain unpleasant feelings. These feelings will reoccur, and so an infant will recognize and reidentify them over TIME via concepts like SAME and DIFFERENT.

Infants will thus acquire concepts such as HURT, which is surely a central feature of PAIN.

And just as concepts of outer objects can be acquired via implicit learning, so too can mental concepts. The infant need not be consciously deploying concepts or consciously acquiring mental concepts. Rather, infants can implicitly recognize similar and central features of recurring mental states and thus reidentify mental states. Infants are thus also able to distinguish one mental state from another and to form rudimentary thoughts using those mental state concepts. These are all key aspects of CONPOSS. Concepts like CAUSE, EFFECT, SAME, THIS, and TIME all play a role from the earliest stages of infancy (if not before). I suggest that once we allow that infants possess concepts, we have little reason to resist the idea they can combine concepts into thoughts (or other intentional states), even if there are some limitations on this ability compared to normal adults (such as inferior inferential capacities). After all, the idea that thoughts are constituted by concepts is rather uncontroversial. To be sure, it does not *logically* follow from this that concepts *must* combine into thoughts, but that is just what concepts do.

Perhaps the main source of resistance to this overall line of argument comes from those "theory-theorists" who hold that infants are not capable of having a "theory of mind" until at least three years. Much of the recent discussion on infant (and even adult) concept acquisition and possession revolves around testing for a so-called mind-reading ability, that is, the capacity to attribute mental states to others or even to oneself. The contemporary literature contains significant work on so-called theories of mind (Carruthers and Smith 1996; Nichols and Stich 2003; Goldman 2006). One of the disputes centers on those who think that our concepts of the mental are acquired through a process of simulating another's mental activity with one's own, and those who argue that some kind of background theory of mind is presupposed in the very ability to mind-read. Thus we have the much-discussed choice between so-called simulation theory and theory-theory of mind, though many authors really hold some form of hybrid view. A related controversy is what exactly the relationship is between understanding one's own mental states (metacognition) and the ability to attribute mental states to others (mindreading). That is, can one be aware of one's own mental states, and thus have mental concepts, without being able to mind-read others at all? Infants and animals are often tested experimentally for their ability to understand whether or not another organism has a perception or belief. Much of the evidence already presented in this

section, such as joint attention studies and evidence for AGENT, strongly suggests that infants have at least some mind-reading ability.

I will return to this overall issue in the next chapter, but will focus here on the so-called false-belief task (Wimmer and Perner 1983), which theory-theorists often use as evidence for their view that infants cannot mind-read. One case goes as follows: Subjects A and B are shown the location of an object, such as a piece of chocolate. It is then moved to another location while subject A is in the room but subject B is out of the room. When B returns, A might be asked where B will look for the chocolate or other behavioral evidence (e.g., expressions of surprise) might be used to determine whether or not A can successfully contrast its own belief from B's belief about where the chocolate will be. Infants do not perform well on these tasks until at least age three, suggesting that they do not have the concept BELIEF or at least cannot distinguish between true and false beliefs in interpreting the behavior or others. Infants frequently respond that B will look for the chocolate in the *new* location, not where B has originally seen it.

However, to my mind, there are a number of persuasive replies to this line of argument against infant mind-reading, including a growing body of counterevidence:

(1) It is first crucial to separate BELIEF, not to mention FALSE BELIEF, from other, arguably simpler mental-state concepts, such as PAIN, PERCEPTION (or SEEING), and DESIRE. Beliefs seem more sophisticated than other mental states and thus make a poor measure of a child's ability to have any mental concepts or any "theory of mind." Concepts of various emotions, such as FEAR or HAPPY, also appear simpler than BELIEF and are more closely associated with certain facial expressions.

(2) Similarly, Zahavi (2004b) questions whether or not it is plausible to suppose that infants who fail the false-belief task still have a lesser understanding of BELIEF. On the surface at least, it seems possible for a child to understand BELIEF but not FALSE BELIEF, which involves a more theoretical understanding of BELIEF. If we allow for degrees of concept possession, then this may also be a way for young children to be credited with the concept BELIEF without fully understanding the possibility of error or the concept FALSE BELIEF. Much as I have argued earlier, Zahavi also presents evidence that infants possess some form of self-awareness and I-concept.

(3) Some have rightly argued that the false-belief task relies too heavily on language and so only measures a more sophisticated theory of mind or understanding of BELIEF (Rochat 2001, 163–165). After all, the children are often asked questions that require a certain level of linguistic competence. It remains likely, however, that young children and infants possess

an implicit understanding of others as intentional agents, at least to some degree. We have already seen significant evidence in favor of this claim.[22]

(4) Bloom and German (2000) argue that the false-belief task should not even be a test for theory of mind. For one thing, it is a particularly difficult task in that a child has to follow the actions of two characters, has to remember both where the chocolate was and is, and has to understand and respond to a question.

(5) Perhaps most interestingly, Birch and Bloom (2004) offer compelling evidence that the main reason for infants' and young children's failure on false-belief tasks has to do mainly with what they call the "curse of knowledge," that is, a "tendency to be biased by one's own knowledge when attempting to appreciate a more naive or uninformed perspective" (256). The idea is that infants overestimate, for example, where B will look for the chocolate due to the curse of knowledge, not because of any conceptual deficiencies. Part of the rationale for this claim is that even *adults* who know the solution to a problem or the outcome to an event tend to overestimate how easy it is for someone else to solve. Adults thus behave in parallel ways on similar tasks. Birch and Bloom point out that this tendency is found in various other contexts, where it is called the "hindsight bias," "the knew-it-all-along effect," and "adult egocentrism," among other names. So if adults tend to behave this way, it would be difficult to hold infants to a higher standard in the false-belief task. Failure in false-belief tasks, then, may have more to do with our natural tendency to project our knowledge onto others in these kinds of tasks. In short, adults and children fail these tasks mainly because they "have a hard time putting aside their own knowledge . . . not because they are unable to appreciate that others can have different perspectives" (Birch and Bloom 2004, 257). They also suggest that a possible underlying explanation is that it is "harder to inhibit one's knowledge than to inhibit one's ignorance" (258). They conclude that "younger children's heightened susceptibility to the curse of knowledge explains their tendencies to overestimate what others know" (259).

(6) Song and Baillargeon (2008) have shown that infants as young as fourteen months are capable of understanding that others form false *perceptions*, which are closely related to *beliefs* in that perceptions are often the basis for beliefs. Infants first watched events in which an agent faced a stuffed skunk and a doll with blue pigtails. The agent preferred the doll, as was indicated by the agent's repeatedly reaching for it. Then, while the agent was absent, the doll was hidden in a plain box, and the skunk was hidden in a box with a tuft of blue hair protruding from under its lid. Using the violation-of-expectation method, infants expected the agent to be

misled by the tuft's resemblance to the doll's hair, and to falsely perceive it as belonging to the doll. Infants can thus keep two different versions of an object in mind, one based in reality and the other corresponding to an agent's false perception (and thus belief).

(7) Buttelmann, Carpenter, and Tomasello (2009) show that false belief understanding is present in most 18 month old infants via a different methodology based on an active behavioral response: helping. So, for example, an infant knows how to open boxes and sees a toy transferred from box1 to box2. Infants choose to help another person by opening box2 in a false belief scenario when the other person does not know that the toy has been move to box2.

In any case, Zahavi (2004b) correctly notes that some theory-theorists tend to adopt some version of HOT theory of consciousness, and some do so explicitly (Perner and Dienes 2003; Carruthers 1996, 2000). This, in turn, leads them to go so far as to cast doubt on infant consciousness generally because infants are allegedly incapable of having mental concepts. However, as we have shown, a good case can be made that HOT theory is indeed consistent with infant consciousness. Mental concepts, or at least a sufficient number of rudimentary mental concepts, can be possessed by infants. Furthermore, it is incorrect to suppose that infant consciousness must be accompanied by *introspective* states (that is, conscious HOTs). Some very coarse-grained mental concepts in unconscious HOTs are enough to do the trick.

Overall, then, I conclude that the Acquisition and Infants Theses are true. I also conclude that they are consistent with the Conceptualism and HOT Theses. A solution to the Consciousness Paradox appears to be within reach. I now turn to the Animals Thesis.

8 Animal Consciousness

In this chapter, I defend the Animals Thesis, which says that most animals are conscious. I also focus mainly on how to reconcile the Animals Thesis with the HOT and Conceptualism Theses, especially since the Animals Thesis, like the Infants Thesis, is widely held. However, some think that animals do not have any concepts. Others have argued that they could not have the sophisticated concepts apparently required by HOT theory. As we saw in the previous chapter, a similar problem faced the Infants Thesis.

In section 8.1, I review and elaborate on a previous exchange between myself and Peter Carruthers who accepts the conclusion that HOT theory entails that most animals are not conscious. I disagree with him. In section 8.2, I argue at length that many animals do in fact have the concepts necessary for HOTs, including both mental concepts and self-concepts. In section 8.3, I critically examine what has come to be known as "Lloyd Morgan's Canon" and argue that attributing conscious mental states to animals is really, in the end, the more parsimonious hypothesis. Since some of the same issues arise for the autistic, in section 8.4 I defend the view that people with autism also have HOTs and mental concepts. Finally, in section 8.5, I argue that conceptualism is consistent with animal consciousness.

8.1 Carruthers, Animals, and HOT Theory

8.1.1 Some Background

The most controversial aspect of Carruthers's views concerns his position on animal consciousness (Carruthers 1989, 1998, 2000, 2005). I have had my say in print on Carruthers's contention that animal consciousness is very unlikely given the truth of some form of HO theory (Gennaro 1993, 1996, 2004b, 2006b). I will not repeat all my arguments here. However, a brief summary is in order, since it can serve nicely as background for the remainder of the chapter.

Carruthers (2000, 195) had at one time presented the following summary of my previous position as follows: "In order for [mental state] M to count as phenomenally conscious, one doesn't have to be capable of entertaining a thought about M *qua* M. It might be enough, [Gennaro] thinks, if one were capable of thinking of M as *distinct from* some other state N" (cf. Carruthers 2005, 49). Carruthers then offered the following reply:

> What would be required in order for a creature to think, of an experience of green, that it is distinct from a concurrent experience of red? . . . Something must make it the case that the relevant *this* and *that* are color experiences as opposed to just colors. What could this be? There would seem to be just two possibilities. [1] Either . . . the *this* and *that* are picked out as experiences by virtue of the subject deploying . . . a concept of *experience*, or some narrower equivalent. . . . On the other hand, [2] the subject's indexical thought about their experience might be grounded in a nonconceptual *discrimination* of that experience as such. (2000, 195; cf. 2005, 50)

Although Carruthers rejects both possibilities, I argued that neither reply is persuasive (Gennaro 2004b). He rejects possibility 1 mainly because "this first option just returns us to the view that HOTs (and so phenomenal consciousness) require possession of concepts which it would be implausible to ascribe to most species of animal" (2005, 50). But I believe that Carruthers overestimates the sophistication of such concepts and underestimates the conceptual capacities of most animals, as I will further argue later in this chapter. For example, he mentions concepts such as EXPERIENCE, SENSATION, and SEEMING RED. But why couldn't animal HOTs contain more modest concepts like LOOKING RED or SEEING RED? Is it so implausible to ascribe *these* concepts to most animals? I don't think so. Animals need not have the concept of "the *experience* of red" as opposed to "seeing or looking red." "I am now seeing red" is a perfectly good HOT. Similarly, even if animals do not have HOTs containing EXPERIENCE in any sophisticated sense of the term, why couldn't they have, say, FEELING?

To use another example, perhaps animals do not have a sophisticated concept of DESIRE, but why not some grasp of the integral notion WANTING FOOD? Once again, perhaps most animals cannot have HOTs directed at pains qua pains, but why can't those HOTs contain THIS HURT or THIS UNPLEASANT FEELING? Having such concepts will then also serve, in the animal's mind, to distinguish those conscious states from others and to reidentify those same types of mental states on different occasions. According to CONPOSS, recognizing M to some degree via a central feature of M and distinguishing M's from non-M's goes a long way toward possessing the requisite concept M. We have already seen analogous considerations supporting the presence of infant HOTs.

Moreover, recall that Carruthers champions the view that there are purely phenomenal concepts of experience, in part, to disarm the explanatory gap argument against reductive materialism. Carruthers acknowledges that the relevant thoughts with recognitional concepts do not have to contain EXPERIENCE: they are those concepts "we either have, or can form . . . that lack any conceptual connections with other concepts of ours, whether physical, functional, or intentional. I can, as it were, just recognize a given type of experience as *this* each time it occurs, where my concept *this* lacks any conceptual connections with any other concepts of mine—even the concept *experience*" (2005, 67). One might therefore wonder why *we* can have such stripped-down demonstrative concepts but animals cannot. Why do animals need to have EXPERIENCE in their HOTs (making it less likely that they are conscious creatures), but *we* don't need to have such sophisticated HOTs? Indeed, the presence of something like recognitional concepts seems precisely to be what Carruthers should allow in response to his first possibility.

Carruthers then rejects possibility 2 mainly because "this second option would move us, in effect, to a *higher-order experience* (HOE [= HOP]) account of phenomenal consciousness" (2005, 50). I had argued (in Gennaro 1996, 95–101) that the difference between the HOT and HOP models is greatly exaggerated. Contrary to what Carruthers says (2005, 50n18), however, I never argued that there is *no* real difference between HOT and HOP theory. Part of my objection did rest on my conceptualist tendencies, so I am skeptical that there are HOPs with analog content, let alone with entirely nonconceptual content. Thus Carruthers's criticism that my view might eventually "move us" to the HOP model is not as damaging as he seems to think. If anything, it seems to make Carruthers's own view *more* likely that animals are phenomenally conscious. HOP theory is normally seen as not having as great a problem in accounting for animal consciousness precisely because the HO perceptual state allegedly is (at least) partly nonconceptual. So Carruthers currently holds a form of HO theory that is normally even friendlier to animal consciousness. He blurs the distinction between the HOP and HOT models by arguing that his dispositional HOT theory is a form of HOP theory (Carruthers 2004). So it is also difficult to see why, in his own view, any move toward HOP would be problematic. In the present book, I have much more explicitly rejected HOP theory (and dispositional HOT theory) in chapter 3 and rejected nonconceptual content in chapter 6.

Moreover, in previous replies to Carruthers, I was careful not to rely solely on the conceptual considerations he cites. I also put forth behavioral, evolutionary, and comparative brain structure evidence for the conclusion

that most animals are conscious. For example, I explained that many animals even have some kind of cortex (Gennaro 1996, 91–95), not to mention the fact that they share with us many "lower" brain structures often associated with conscious states in humans. Carruthers's failure to put our disagreement in context is significant because the cumulative effect of such strong inductive evidence in favor of animal consciousness is lost.[1]

Now why exactly does Carruthers think that most animals don't have HOTs? The primary reason has to do with his allegiance to the "theory of mind" theory, whereby understanding mentalistic notions presupposes having a "folk-psychological" theory of mind. Once again, however, Carruthers builds a great deal into having such a theory and explicitly ties it to the capacity to have HOTs. For example, he cites experimental work by Povinelli (2000) and others suggesting that chimps lack APPEARING or SEEMING or PERCEPTION (as subjective states of the perceiver), which he takes as necessary to have HOTs about experiences. Such experiments are often designed to determine if chimps notice whether or not the experimenter is looking at or away from something (such as food). In line with many theory-theorists, Carruthers holds that animals with HOTs should be able to have thoughts about the mental states of *other creatures*, as, for example, we might expect to find when (or if) animals engage in deceptive behavior.

I tackle this issue at length in the next section. I disagree that one should conclude from such evidence that (most) animals don't have HOTs. I will make two brief points here. First, it is not clear that we should read too much into the failure of animals in such experiments. For one thing, these are obviously not the natural conditions or environments of the animals in question. Second, even if some or most animals cannot, say, engage in deceptive behavior and so arguably do not have HOTs about the mental states *of others*, it still does not seem to follow that they cannot have less-sophisticated HOTs about *their own* mental states. After all, unconscious self-directed HOTs are all that are required for conscious states, according to HOT theory.

8.1.2 Unconscious Suffering and Frustration?

With an eye toward linking his view of animal consciousness to moral issues, Carruthers had previously argued that creatures with only unconscious pains—pains that would lack any subjective qualities, or *feel*—could not be appropriate objects of sympathy and moral concern. But Carruthers has changed his mind. In Carruthers 2005, chapters 9 and 10, he is concerned to show that animals can be objects of sympathy and moral concern because the "most basic form of mental . . . harm lies in the existence

of thwarted agency, or thwarted desire, rather than in anything phenomenological" (157). Indeed, according to Carruthers, frustration, suffering, grief, and disappointment can all occur in the absence of phenomenal consciousness because many organisms, still capable of having (unconscious) mental states, can find themselves in a situation such that there is "the co-activation within a creature's practical reasoning system of a first-order desire together with the first-order belief that the state of affairs that is the object of the desire doesn't obtain" (177). For example, suppose that an animal currently wants to drink but believes that it isn't presently drinking. Animals can still be averse to unconscious pains in the sense that they take steps to avoid being in that state. Carruthers interestingly argues in various places (2005, chap. 12; 2009a) for the view that it is perfectly reasonable to attribute all kinds of (unconscious) intentional mental states, and thus concepts, to animals, even ants and bees.

I still find Carruthers's arguments for his current moral stance unconvincing. For example, he asks us to imagine a conscious, language-using agent called Phenumb, "who is unusual only in that satisfactions and frustrations of his conscious desires take place without the normal sorts of distinctive phenomenology" (2005, 172). I will not get bogged down in the details of Carruthers's thought experiment here, but he ultimately argues that Phenumb is an appropriate object of moral concern and that the example shows "that the psychological harmfulness of desire-frustration has nothing (or not much) to do with phenomenology, and everything (or almost everything) to do with thwarted agency" (173). In essence, Carruthers is attempting to separate desire frustration from consciousness to make room for the idea that unconscious animals can be the objects of sympathy and moral concern, contrary to his previously held position.

I am puzzled by Carruthers's argument for several reasons. First, the hypothetical Phenumb begins as a conscious agent, and it seems to me that desire frustration is a more sophisticated intellectual psychological capacity than the mere ability to subjectively *feel* pains. Even if the two capacities are somehow *theoretically* distinct, I fail to see what positive reason we could ever have to attribute *only* the former to any known animal. Second, even if we can imagine the possibility of this Spock-like character only able to have such purely intellectual frustrations (as Carruthers suggests in 2005, 172n15), it does not follow that such frustrations would be entirely nonphenomenal. Carruthers is curiously comparing (what he takes to be) unconscious animals to a highly sophisticated intellectual hypothetical character.

Third, when Carruthers speaks of "desire frustrations," it is unclear how they could all be unconscious. I am not even sure I understand the idea of a *non*phenomenal "disappointment" or "desire frustration." Of course there can be unconscious desires (and even, I think, unconscious pains), but it does not follow that there are unconscious desire *frustrations*, especially in organisms who are supposed to be utterly unconscious. Thus, in the end, I don't believe that Carruthers's current moral stance is any more tenable than his previous view. No doubt part of the problem is terminological. One can, I suppose, *speak* of unconscious "sufferings," "feelings," and "desire frustrations," and we are all entitled to use our own terminology to some extent. However, as we saw in chapter 1, there is a point where using terms in this way becomes more of a provocative attempt to redefine them and simply adds to the terminological confusion. It is most important, though, to keep our sights set on the issue of whether or not actual animals have conscious mental states.

Carruthers might at this point accuse me of implicitly holding a Searlean position such that each and every unconscious mental state is either actually or potentially conscious. But I hold no such view and reject Searle's Connection Principle (see chapter 2). There is a middle ground ignored by Carruthers, namely, that some kinds of mental states (pains, frustrations, sufferings) can only be had by conscious *organisms*. That is, from the fact that *we* have unconscious pains and feelings, it doesn't follow that there are (or could be) organisms with *all* unconscious pains and feelings.[2]

Another way to think about this is by comparing animal behavior to some current-day robots. If we are convinced that a robot is utterly unconscious, we may still (rightly, I think) attribute to it beliefs, desires, and perhaps even perceptions if its behavior is complex enough. However, it is not clear that the same would or should go for pains, sufferings, frustrations, and disappointments. These are arguably parasitic on the *creature* or *system* being conscious in the first place. It seems to me that we wouldn't (and shouldn't) ever say that a robot is suffering unless we were convinced that it is capable of otherwise having phenomenally conscious states. Indeed, we are often inclined to think in terms of the entire organism as a conscious agent when attributing such states as frustrations or disappointments: "I am frustrated," "The dog is suffering," and "My sister is very disappointed." But this is not to endorse either the stronger Cartesian view or even the weaker Searlean Connection Principle. It is consistent with holding that any individual mental state *in a conscious creature* can be unconscious. The key difference, I think, lies in the fact that beliefs and desires

are best understood purely as dispositions to behave in certain ways and thus are more reasonably attributed to utterly unconscious robots or even to unconscious insects.

Carruthers, however, does raise the important issue of the relationship between intentional states (such as beliefs and desires) and consciousness. He clearly also treats desires and beliefs as somewhat more primitive mental states, which can even be attributed to ants and bees (2005, chap. 12). In these cases, the behavior in question is at least arguably complex enough to warrant such mental ascriptions. It is a notch up from the purely inflexible fixed-action behavior patterns found in some primitive insects.

It is also unlikely that the common belief that many animals have conscious states (which in turn cause their behavior) can be so easily explained away as an anthropomorphic process of "imaginatively projecting" what it is for us to have certain mental states onto various animals (as Carruthers argues in response to Lurz on 198–200). Carruthers insists that we are under an illusion in thinking that phenomenal consciousness is needed to explain the cause of any animal behavior. He calls it the "in virtue of illusion" whereby we mistakenly think that it is in virtue of the phenomenally conscious properties of experience that we behave the way we do. If someone picks out a tomato by its color, one will normally have a phenomenally conscious experience of red, but it does not follow that the phenomenal property *causes* the behavior in question. For Carruthers, then, although a human phenomenally conscious *state* does cause the behavior in question, it is the first-order (not higher-order) *content* of a conscious state that does virtually all the causal work. Carruthers contends that we can give a similar explanation for a variety of animal behaviors, except that they are not phenomenally conscious in the first place (that is, there is no higher-order content to the state at all).

This move by Carruthers strikes me as highly implausible and certainly does not seem to describe what I am thinking when I attribute conscious mental states to animals. I agree with Lurz that the central initial reason for believing that animals have conscious mental states has more to do with the fact that their rather complex behavior is best explained and predicted by attributing such folk psychological notions to them (cf. Saidel 2009; DeGrazia 2009). We *could* of course be wrong and under a massive illusion here, but I am not convinced that we are. As Carruthers recognizes, he is thus dangerously close to embracing some form of epiphenomenalism whereby conscious states (or at least the property of consciousness) have no causal impact on one's behavior (Carruthers 2005, 186–187, 204–206). But in some places, he does not wish to go as far and says that "phenomenal

consciousness might be *almost* epiphenomenal in its functioning within human cognition" (195).

Carruthers has much to say about how animal behavior can be explained without appeal to consciousness, but then one even begins to wonder why we are forced to attribute consciousness to other humans. He relies on the aforementioned two-visual-systems hypothesis and cites familiar evidence from blindsight cases to show how some surprising behavior can occur unconsciously (2005, 204–206). However, he overlooks the important fact that blindsight patients do not *voluntarily* act toward objects in their blind fields. They only act in response to forced guesses in response to the examiner. So it does not seem to me that blindsight cases support Carruthers's view, at least to the extent that animals behave voluntarily or "on their own" toward objects of perception. We do indeed seem to behave, at least sometimes, in virtue of our conscious perceptions.

For example, an unexpected sound behind an animal may cause it to flee. The animal is not being forced to guess whether or not something is behind it, as an analogue of blindsight would require. The best explanation still seems to be that it fled because it consciously heard the sound and feared for its life, at least for animals with complex-enough behavior and similar comparative brain areas. How could unconscious visual experiences (and smells and sounds) cause a seeing-eye dog to help its owner cross a busy street? After all, the owner needs the dog precisely because she has lost her *conscious* vision. Similarly, navigating through the world by means of blindsight is obviously not good enough, not to mention extremely dangerous. One can only imagine how much worse off one would be if one also lost all conscious sense of touch or smell or hearing. But if Carruthers is right, having an unconscious dog is somehow able to help a blind person. Moreover, if *we* have two visual systems (with one entirely unconscious and the other conscious), then it is surely reasonable to hold that many animals, whose brain structures are similar, also have the conscious system. As I understand it, many animals do indeed have both visual systems. Carruthers might again insist that those animals do not have a "HOT faculty." But then why would evolution have produced *two* visual systems in so many animals if they are both unconscious?

Despite Carruthers's insistence that he has "no axe to grind" (2005, 181), it is difficult not to notice the elaborate attempts to explain away the plethora of evidence for animal consciousness (and HOTs) as misleading, mistaken, or illusory while any piece of evidence suggesting the presence of (unconscious) intentional states is interpreted in the most favorable light. To be sure, Carruthers presents interesting reasons and important

arguments in support of his position. And it is perfectly fine to push one's line of argument as far as it goes, but his is a puzzling combination of views, in my opinion. Moreover, it is surprising that Carruthers would say that there "is no radical Cartesian divide here, between genuinely minded humans and mere mindless automatons" (2005, 204). This may be true if we were talking only about unconscious mental states reaching far down the evolutionary chain, but the "great Cartesian divide" has always had much more to do with any alleged radical difference between human and animal *consciousness*. That is the real Cartesian divide. After all, Descartes didn't even believe in unconscious intentional states. It is clear to me that humans would (and should) treat animals differently if we were convinced that they were not conscious, contrary to Carruthers's claim that "very little of significance for comparative psychology need follow from the fact that phenomenal consciousness in denied to many non-human animals by higher-order thought theories of consciousness" (204).

Finally, given his views about animal consciousness, one wonders just what entitles Carruthers to appeal repeatedly to "common sense" and "authority" when it otherwise suits his purposes (2005, 216–217). Why should anyone accept what "most people" think about, say, ascribing beliefs and desires to animals from someone who rejects similar logic regarding animal consciousness?

It is worth mentioning at this point that FO theorists also need to account for animal consciousness. After all, even if one ties consciousness to first-order intentional states (such as beliefs, thoughts, and desires), the question still remains as to what organisms are capable of having these states. Don't they require some concept possession too? Although Tye agrees that honeybees and fish are conscious (Tye 2000, chap. 8), he argues that *suffering* "requires the cognitive *awareness of pain*" (182; italics mine), which these presumably simple minds lack. However, Tye speaks of animals' needing "the power to introspect" (182) in order to suffer, which he thinks many animals do not have. But lacking introspection would not even rule out animal suffering according to HOT theory, since only unconscious HOTs are needed for animal pain and suffering. I am not sure, for example, if bees or other insects are conscious or suffer, but we need not require them to have *conscious* HOTs.[3]

8.2 Animals and I-Thoughts

Following up on the dialectic set out in the previous section, then, let us recall that I-thoughts are thoughts about one's own mental states or about

"oneself" in some sense. Whether or not animals have I-thoughts has become a central topic of empirical investigation. As we have seen, I-thoughts are closely linked to what psychologists call "metacognition," that is, mental states about mental states, or "cognitions" about other mental representations (Bennett 1988; Metcalfe and Shimamura 1994; Koriat 2007). Although some reject the notion that most nonhuman animals have I-thoughts, the evidence seems to be growing that many animals are capable of having I-thoughts and have some ability to understand the mental states of others (Terrace and Metcalfe 2005; Hurley and Nudds 2006; DeGrazia 2009). Of course, a HOT (or MET) is a kind of metacognitive or metapsychological state, which is of the form "I am in mental state M now." The allegation, however, is that HOT theory rules out animal consciousness because animals (or at least most animals) do not possess such sophisticated I-concepts and mental concepts.

This objection to HOT theory is normally presented by non-HOT theorists, such as Dretske (1995), Lurz (2002, 2004), and Seager (2004). As we have seen, however, one prominent HOT theorist, Peter Carruthers, embraces this alleged consequence of HOT theory. Since most of us believe that many animals have conscious mental states, a HOT theorist must explain how animals can have the HOTs necessary for such states. Once again, a reason that most of us naturally believe that animals have conscious states is simply that our folk psychology is a theory of conscious mental states, and it works well in explaining and predicting much of animal behavior.[4]

Thus there is a three-way tension among the following claims that needs to be relieved:

(a) Most animals have conscious mental states; that is, there are generally positive grounds for believing that animals have conscious states independently of any commitment to a philosophical theory;
(b) the HOT theory is true, which, in turn, entails having I-thoughts; and
(c) few (if any) animals are capable of having I-thoughts based on various empirical and theoretical considerations.

Carruthers rejects (a) and embraces (b) and (c), while Dretske, Lurz, and Seager endorse (a) and (c) but reject (b). I reject (c) and accept (a) and (b). Thus this section has a double purpose: to discuss and elaborate on the evidence for HOTs (or I-thoughts) in animals and to show that the HOT theory is indeed consistent with animal consciousness. The Animals Thesis is both true and consistent with the HOT Thesis. I argue that recent experimental evidence on animal memory and metacognition strongly suggests that many animals have the self-concepts and mental-state concepts necessary to form

I-thoughts. I also reply to the claim that having I-thoughts requires having thoughts (and thus concepts) directed at *others'* mental states.

The stakes are therefore extremely high, because if HOT theory is true, any evidence indicating the absence of I-thoughts would also serve to cast doubt on animal consciousness itself. Of course, most researchers in the field do not question that animals have at least some basic conscious states, such as perceptions, pains, and desires. Nonetheless investigators should be aware of this potential consequence in some philosophical circles.

It is also crucial to remember throughout this chapter that when a conscious mental state is a first-order world-directed state, the higher-order thought (HOT, or MET) is not itself conscious. When the HOT is itself conscious, there is a yet higher-order (or third-order) thought directed at the second-order state. In this case, we have introspection.

It is sometimes said that all or most (nonhuman) animals cannot mind-read; that is, they do not understand that others (or even they) have mental states. In addition, adherence to the so-called theory-theory view of mind reading, whereby understanding mentalistic notions presupposes having a "folk-psychological" theory of mind, seems to rule out that animals have I-thoughts.[5]

As I mentioned earlier, Carruthers cites experimental work suggesting that chimps lack APPEAR or SEE (Povinelli 2000), which he then treats as necessary for HOTs about *one's own* experiences. Such experiments are often designed to determine if chimps take notice of whether or not the experimenter is looking at something (say, food) or is unable to see something (for example, due to blindfolding). Carruthers argues that animals with HOTs should also be able to have thoughts about the mental states of *other creatures*. However, the evidence seems to be growing that many animals can indeed have I-thoughts and mind-read, and it is not clear that having I-thoughts requires reading *other* minds.

So there are two main concepts in an I-thought or HOT, namely, a self-concept ("I") and mental-state concept ("M"). Let us consider them in turn.

8.2.1 Self-Concepts and Episodic Memory in Animals

Recall that episodic memory (EM) is a personal and explicitly conscious kind of remembering involving "mental time travel" (Tulving 1983, 1993, 2005). It is often contrasted with *semantic* memory, which need only involve knowing that a given fact is true or what a particular object is, and *procedural* memory, whereby memory of various learned skills is retained. Tulving also uses the term "autonoetic consciousness" or "autonoesis" for the kind of consciousness characterized by episodic remembering. The link

to I-thoughts is fairly clear: some notion of self or "I" seems necessary to have a genuine EM. I recognize the EM as *mine* and representing an event in *my* past. Tulving speaks of EM's "dependence on a remembering 'self' [and] . . . relation to subjectively apprehended time" (2005, 14).

However, Tulving himself resists the idea that nonhuman animals have EM. But his case is less than convincing, and he even seems to retreat from this position later in the same essay. For example, he concedes "that some of [his] assertions may be too strong" (2005, 27). In speaking, by analogy, of amnesic patient KC, he explains that "it is possible that he [KC] has a little left of one or more of the properties of episodic memory and that we are dealing with a case of severe impairment in episodic memory rather than its total absence" (27–28). This naturally makes one wonder why the same should not be said for most nonhuman animals if indeed KC has at least a limited ability for EMs (say, into the more immediate past), as well as an ability to think about the short-term future. It is hard to see how, for example, KC can play various card games and even chess without having *some* ability for mental time travel (both into the past and future). Chess surely involves planning one's moves and remembering certain strategies. Tulving also tells us that KC's short-term or "working memory" is "normal; he remembers what happened a short-term while (1 to 2 minutes) ago" (23).[6]

Turning to nonhuman animals, Tulving often qualifies his strong negative claim when speaking approvingly of the more cautious conclusion (reached in W. Roberts 2002) that chimps' ability to imagine their *extended* future is in doubt and that their ability to mentally travel into the future and past is *limited* (39). Perhaps most telling is Tulving's concession that Clayton and Dickinson and their colleagues (in Clayton, Bussey, and Dickinson 2003) have "reported ingenious and convincing demonstrations of memory for time in scrub jays" (37). Scrub jays are food-caching birds, and when they have food they cannot eat, they hide it and recover it later. Because some of the food is preferred but perishable (such as crickets), it must be eaten within a few days, while other food (such as nuts) is less preferred but does not perish as quickly. In cleverly designed experiments using these facts, scrub jays are shown, even days after caching, to know not only *what* kind of food was *where* but also *when* they had cached it (see also Clayton, Emery, and Dickinson 2006).

We also have much more recent evidence for EMs in various animals, not mentioned at all by Tulving. Evidence in primates is discussed in the same volume in which Tulving's 2005 chapter appears. Menzel (2005), for example, describes experiments in which a female chimp (Panzee) recovers

food hidden by a trainer by coaxing a different trainer (who was unaware of its location) to let her get it. This was done even after quite a bit of time delay between the observation and the time at which the second trainer appeared. Schwartz (2005) presents a similar result with respect to a gorilla named King. And Hampton (2005) found that a monkey was able to successfully match-to-sample even after delays. The subject had to decide whether to submit to a test of the sample *after* the stimulus display had been removed but *before* the test has been presented. The results indicate that monkeys both know when they remember and when they have forgotten, indicating both a capacity for EM and a form of metacognition.

Additional evidence for EMs in animals is presented in Eichenbaum et al. 2005. For example, rats can remember the temporal order of the odors of various objects. And Dere et al. (2006) extensively review the expanding literature strongly suggesting that EMs exist in various other animals, including dolphins, birds, and rodents (such as mice and rats). Finally, it is interesting to note that there are also reports of animals' ability to plan for the future, such as is found in western scrub jays (Raby et al. 2007). They show that the jays plan for future need by preferentially caching food in a place where they have learned that they will be hungry the following morning. It is often observed how an ability to think about the future ("prospective memory" or "prospective cognition") is closely related to the capacity for EM (a point Tulving repeatedly stresses in his 2005 chapter).[7]

Some interpretations of the above data are, to be sure, not uncontroversial in some circles, and we should not jump to unwarranted conclusions. However, there is little reason to hold the very strong view that animals have absolutely *no* EMs or no ability at all to "mentally time travel." If this is correct, then there is also no reason to deny an animal's ability to form at least some minimal self-concept, which, in turn, can figure into a HOT.[8]

Three other points are worth making here. First, it is important not to equate having EMs with having *accurate* EMs. That is, if an experiment really shows a lack of the accuracy of an EM, we should not conclude that the animal in question has no EMs at all. If we compare this to human adults, we can quickly see the problematic inference. Human eyewitness reports, for example, are often mistaken, but we would not and should not infer any general lack of EMs on the part of the subject. We have all had students with an almost inexplicable inability to do well on certain exams; yet they surely have EMs otherwise. The same person can often be much better at remembering certain things (such as song lyrics or baseball statistics) and not others, even if we think the others are more important. The same surely

goes for various past experiences. We must be careful not to hold animals to a higher standard than we hold ourselves. Human memory routinely fails in various experimental and everyday contexts.

Second, I have previously argued at length that there is a compelling a priori Kantian-style argument showing that having at least some form of EM is necessary for being a conscious creature (Gennaro 1992; 1996, chap. 9). The basic idea is that having concepts of outer objects involves understanding those objects as enduring through time (since we do not take them to be mere fleeting subjective states of mind), which, in turn, requires us to think of ourselves as temporally enduring subjects with a past (mainly because we recognize that those objects are the same objects at different times). That is, if a conscious organism can reidentify the same object at different times, then it implicitly understands itself as something that endures through time. We also saw this theme at work in the previous chapter with respect to infant consciousness and early or innate concepts.

Third, recall from chapter 7 that there are degrees of self-concepts. One also finds a willingness to talk of a continuum of self-consciousness in the animal cognition literature (Kinsbourne 2005), as well as corresponding levels of consciousness in human development, including the purely physical "self-other contrast" (Nelson 2005). In any case, all that is needed for having most HOTs is the kind of minimal "bodily self-consciousness" self-concept, that is, being able to distinguish one's own body from other things (DeGrazia 2009). Like infants, it would seem that animals are clearly also capable of at least having some kind of bodily awareness. It is surely fairly uncontroversial that most animals at least have this unsophisticated I-concept. And recall that one defining feature of concept possession (in CONPOSS) involves discriminating between objects or properties. In this case, we at least have an ability to distinguish between oneself and others. In the end, however, I think that many animals are capable of more sophisticated self-concepts, as is evidenced by the results and arguments offered in this and the next section.[9]

Finally, it is worth mentioning that recent evidence and the foregoing argument can also be used against one of the best-known arguments against I-thoughts in most animals. Jonathan Bennett (1964; 1966, 116–117) argued decades ago that animals cannot have past-tense thoughts of the form "I was F in the past." This is because most animals do not have the requisite concept of self and because past-tense thoughts cannot be possessed without language. It seems to me that the evidence cited already seriously calls Bennett's argument into question.

8.2.2 Animals and Mental Concepts

Let us look now at evidence for *mental-state* attributions, that is, the "M" part of a HOT. Some of the evidence suggests that animals have metacognitive states, but only directed at themselves, thus, I-thoughts. We have already seen some evidence for this in the previous section in what is sometimes termed "metamemory." If an animal has a metamemory state, then it not only has a self-concept but also is able to form a thought directed at a memory, which is itself a mental state.

In addition, there is the much-discussed work on uncertainty monitoring with animals such as monkeys and dolphins (J. Smith, Shields, and Washburn 2003; J. Smith 2005). For example, a dolphin is trained in a perceptual discrimination task, first learning to identify a particular sound at a fixed frequency (the "sample" sound). Later he learns to match other sounds to the sample sound. When presented with a sound that is either the same or different in pitch as the sample sound, he has to respond in one way if it is the same pitch (such as by pressing one paddle) and another way if it is a different pitch (pressing another paddle). Eventually the dolphin is introduced into a test environment by being forced to make extremely difficult discriminations. To test for the capacity to take advantage of his own uncertainty, the dolphin is presented with a third "uncertain" response that is rewarded if he is uncertain. He is presented with a third paddle, the Escape paddle, which is virtually equivalent to declining the trial. The dolphin chooses the Escape paddle with expected frequency and a similar response pattern to humans and rhesus monkeys, which many researchers take to suggest that the dolphin is aware of his state of uncertainty, that is, he has some knowledge of his own mental state. This is a metacognitive state: the dolphin is aware that he doesn't know something, in this case, whether or not a sound matches (or is very close to) the sample sound. It seems reasonable to seek a common underlying explanation for all the subjects involved (Browne 2004).

A related paradigm has to do with a subject's (such as a monkey's) ability not to respond accurately to a stimulus, but rather on the appropriateness of her level of confidence in the accuracy of a response (Son and Kornell 2005; Hampton 2005). Such "metaconfidence judgments" are treated as evidence of metacognition because the experiments are designed to elicit a "betting" judgment of the form "I am confident that I know" or "I am not very confident that I know." I will not describe this experiment in great detail, but, for example, two rhesus macaques were tested in this way using a system of low- and high-risk bets. In brief, the monkeys tended to bet "high risk" much more often when they were able to make accurate confidence

judgments, and bet low risk more often when responding incorrectly. Thus they seemed able to express feelings of confidence or lack of confidence about their cognitions, in this manner displaying a metacognitive ability.

It is also crucial to note here that some authors do speak of degrees of metacognition or self-awareness in ways arguably similar to the distinction between conscious HOTs and unconscious HOTs. For example, Son and Kornell (2005, 300–301) talk of "*implicit* meta-cognition" in ways that sound very much like unconscious HOTs. They refer to the "*tacit* meta-judgment of uncertainty" as opposed to *explicit* metacognitions or *self-reflective* consciousness, which clearly has an affinity to the more sophisticated conscious HOTs. We may not often have evidence of such further awareness (that is, self-reflective consciousness) because monkeys, for example, cannot verbally express their judgments. "If we did, we would then have evidence of *meta*-metacognition" (318). Again, a clear example of meta-metacognition would be a *conscious* HOT, which is more than is required for one to have a conscious state. Recall that only an unconscious HOT is needed for having a first-order conscious state. Thus animals can still have unconscious HOTs continuously accompanying all their conscious states. As Son and Kornell put it: "According to this view, we make meta-cognitive judgments *constantly* and without explicit knowledge of them" (2005, 317; italics mine). Kinsbourne (2005, 152–155) also speaks of degrees of "self-awareness." I agree when he says that "self-awareness is a matter of degree. Its most crystallized state should be considered an end point of a continuum that emerges from infant and animal experience" (153).[10]

Let us turn to the ability of animals to attribute mental states *to others*, which is another kind of HOT, albeit a thought about *another's* mental state. Despite the Povinelli-style experiments briefly noted earlier, the evidence seems to be growing that at least some animals can mind-read under other or more familiar conditions. For example, recent work by Laurie Santos and colleagues shows that rhesus monkeys attribute visual and auditory perceptions to others in more competitive paradigms (Flombaum and Santos 2005; Santos, Nissen, and Ferrugia 2006). Rhesus monkeys preferentially attempted to obtain food silently only in conditions in which silence was relevant to obtaining food undetected. While a human competitor was looking away, monkeys would take grapes from a silent container, thus apparently understanding that hearing leads to knowing on the part of human competitors (Santos, Nissen, and Ferrugia 2006). Subjects reliably picked the container that did not alert the experimenter that a grape was being removed. This suggests that monkeys take into account how auditory information can change the knowledge state of the experimenter. In

addition, Rhesus monkeys also chose to take food from human competitors who could not see them, either because the humans' eyes were facing away or because their faces were blocked by an opaque barrier (Flombaum and Santos 2005). In a similar vein, it has also been argued that many animals' ability to live complex social lives and to take into account another's spatial perspective provides further evidence for mindreading (DeGrazia 2009).[11]

It is important to point out that controls are used to eliminate at least some non-mind-reading (or "behavior-reading") interpretations of the data. For example, it is shown that monkeys do not prefer to steal grapes from the nonbelled containers simply because the sound of bells frightens them. There was also no historical link between the competitor's observable features and his future actions; that is, subjects had no past experience hearing the bell make noise that could have been associated with the competitor's likely response.

Now, it may be the case that a non-mind-reading interpretation *could* still be given for all or virtually all such experiments. It may indeed *always* be *possible* to construct creative and often elaborate alternative first-order mental or even purely behavioral explanations for any given set of animal behaviors (Povinelli and Vonk 2006; Carruthers 2008). A thorough reply to this line of argument could be a topic for another book, but I will address it more fully in section 8.3. I'll only say here, first, that just because an alternative explanation is *possible*, it doesn't follow that it is the best or most reasonable explanation. Second, there comes a point where such deflationary interpretations might even work for much of *human* behavior or for, say, the behavior of a deaf-mute human who is incapable of verbal communication.

Finally, it is worth mentioning that experimental results similar to those described earlier on caching and episodic memory are applicable here as well. For example, many crows and scrub jays return alone to caches they had hidden in the presence of others and recache them in new places (Emery and Clayton 2001). This suggests that they know that *others* know where the food is cached, and thus, to avoid having their food stolen, they recache the food.

Taken together with the earlier evidence presented, it seems reasonable to suppose that many animals have I-thoughts of some kind or other. Although many subjects of the experiments described earlier are primates, many other "lower" animals are also tested, such as dogs, pigs, dolphins, and even mice and rats. In the next section, I now consider two problematic claims that underlie much of the opposition to my conclusions.

8.2.3 Two Problematic Claims

There are those who insist that if an animal cannot pass a given mind-reading task directed at another, the subject animal is therefore incapable of having any thoughts about its own mental states. This problematic claim is the main target of this and the next subsection. The thesis to be challenged, then, is:

(1) Having I-thoughts requires having thoughts directed at others' mental states.

The obvious corollary of (1) is:

(2) If an organism O cannot form *concepts* of another's mental states, then O also cannot have I-thoughts of any kind.

In reply to thesis (1), we must first acknowledge that despite the evidence presented in the previous subsections, many animals do not pass various tests designed to show the ability to mind-read (such as the Povinelli-style experiments described earlier). Nonetheless I strongly disagree with the notion that one should conclude from these negative results that most animals therefore do not have any HOTs. That is, I believe that (1) is false.

First, it is not at all clear that so much should be read into the failure of animals in such experiments. For one thing, these are obviously not the natural conditions or environments of the animals in question. Perhaps failure can be explained in such situations because they don't typically arise in their native environment. As we have seen, many primates, at the least, do much better in similar tests when performed in more natural or competitive settings. Moreover, it is odd to treat such experimental results as if the paradigms used indicate an undeniably clear *necessary* condition for having other attributing mental capacities. This is somewhat reminiscent of treating failing the Turing Test as indicating the utter absence of any intelligence or even consciousness. Many animals could not pass the Turing Test, but we surely shouldn't conclude *on that basis* that they entirely lack intelligence or are not conscious.

More to the point here, even if some or most animals cannot, say, engage in intentionally deceptive behavior and so arguably do not have thoughts about the mental states *of others*, it still does not seem to follow that they cannot have *unconscious* HOTs about *their own* mental states. Recall that according to HOT theory, only self-directed HOTs are required for conscious states. Thus I agree with Ridge (2001, 333) that the opposing view rests on the false assumption "that there could not be an agent capable of having

HOTs about its own mental states but incapable of having HOTs about the mental states of others." This is a major issue in its own right, namely, to what extent the HOT model requires "mind reading" of *others* as opposed to self-monitoring or metacognition. Nonetheless mind reading of others seems more sophisticated and does not seem necessary for there simply to be conscious mental states, especially simple conscious pains or perceptions. Moreover, as Ridge (2001, 322) also points out, the move from "no deceit" to "no HOTs whatsoever" is much too quick and unjustified. Just because evidence of intentional deception would be the best or clearest evidence for mind reading, it clearly does not follow that the lack of such evidence indicates a lack of HOTs.[12]

Finally, it seems to me that some tests for other-attributing thoughts in the cognitive ethology and theory-of-mind literature are really more often aimed at determining whether or not animals (or infants) can have *conscious* HOTs directed at another's mental state. Speaking of "intentional deception," for example, suggests that the animal is *consciously* intending to cause a false belief in another animal. To the extent that this is what experimenters have in mind, there is again no reason to think the HOT theory is in trouble. As we have seen, the HOT theory allows for the presence of conscious states even in the absence of any (either self-attributing or other-attributing) *conscious* HOTs. Once again, the HOT theory only requires *unconscious* HOTs for first-order conscious states. If a HOT is itself conscious, then one is in a more sophisticated *introspective* state, which is not necessary for having a more primitive first-order conscious state.[13]

Thus thesis (1) is, at best, on very shaky ground.

8.2.4 Concept Possession and the Generality Constraint

One might reply to the foregoing argument that adherence to the so-called generality constraint dictates that if one can self-attribute a mental concept, then one should be able to other-attribute that mental concept. In an effort to show this to be a serious problem for the HOT theory, Seager (2004) also cites experimental evidence suggesting that animals do not other-attribute mental states, and then says the following:

> All animals also lack the ability to attribute mental states to themselves because those who can self-attribute will be a subset of those who can other-attribute. . . . It is . . . [a] doubtful logical possibility that a being which lacked the ability to attribute mental states to others could attribute them to itself. . . . Such an asymmetry would seem to run counter to Evans's (1982) 'generality condition' on concept possession (the claim that one cannot have a concept C unless for any object O, one can have the thought that 'O is C'). . . . What is incoherent is the notion that one could conceptualize one's

own possession of mentalistic attributes while being completely unable to form a thought of another being having mental states. (Seager 2004, 264–265)

Recall that the generality constraint is sometimes put as follows: the attribution of thoughts to any organism of the form "a is F" and "b is G" commits us to the idea that the organism should also be able to think that "a is G" or "b is F." Moreover, "the content of all propositional attitudes is said to be subject to this constraint" (Toribio 2007, 446). Thus we might also think of the generality constraint as involving a commitment to the idea that belief and desire states (which are also composed of concepts) can be recombined with other such states, and perhaps even that the organism can make appropriate simple inferences among them.[14]

In any case, Seager's quotation, at minimum, indicates an endorsement of thesis (2): if an organism O cannot form *concepts* of another's mental states, then O also cannot have I-thoughts of any kind. However, as we have seen, the generality constraint is, by virtually all accounts, a very strong condition to be placed on concept possession. It makes all concept users into idealized rational agents capable of combining into thoughts all concepts that one possesses. Moreover, one might instead opt for the view that possessing a concept C is one of degree, which allows for a partial understanding of C.

Even among human beings, we sometimes distinguish between someone who has a "partial" concept of, say, DOG and someone who has, on the one extreme, "no concept at all" and, on the other, an expert biologist's dog concept. The same might be said for the understanding that a young child has of TREE compared to an adult's increased understanding and then finally to a botanist's tree concept. Treating concept possession in this way would also not require all concept users to be able to draw all possible connections among one's stock of concepts. We sometimes might not "see" the connection between two concepts or thoughts containing those concepts because we have only a partial grasp of the concepts involved. This should also be said for various animals; that is, it may well be that animals have a (partial) understanding of EXPERIENCE, VISION, SEE, PERCEPTION, and the like, without always being able to apply those concepts to others in certain experimental situations. We should not conclude that animals have *no* concept of C if they are unable to pass a test for a more advanced understanding of C. This, I suggest, is what skeptics often in fact do when, say, a chimp doesn't infer a concept of visual experience from certain particular movements (or lack of movements) of another's eyes. Instead we have seen how various experiments (such as Santos's) indicate at least some partial grasp of SEEING and HEARING.

This notion of concept possession, reflected in CONPOSS, can thus help to explain *why* some animals arguably do not make certain connections between thoughts and other propositional attitudes. It can also explain how an animal can fail to attribute a mental concept to another but still apply that concept to itself.

Of course, one central issue then becomes: what is it to have a mental concept M anyway? Again, the answer will in part depend on one's notion of concept possession. If we accept something like CONPOSS along with the idea that one can have a concept C based on a partial understanding of C, then it becomes unclear what positive reason we have for withholding concepts such as PERCEPTION, PAIN, and DESIRE from most animals. To be sure, many animals may not have a sophisticated concept of these mental states, but, given the foregoing experimental results, it seems that they are aware that they are in such states, such as a dog being aware that it is seeing as opposed to hearing. Similarly, an animal knows when it is in pain even if it has a difficult time determining or understanding when another creature is in pain. Thus, once again, having a partial understanding of the concept PAIN might simply involve the notion that "this hurts," as opposed to, say, comprehending philosophical writings on pain. An animal with a DESIRE for food understands that it "wants something." And a partial understanding that one is having a VISUAL PERCEPTION involves at least grasping that one is seeing as opposed to hearing. Recall that Allen makes a similar point: "Philosophers have been tempted by the argument that . . . for example, a dog does not believe there is a squirrel in the tree because it lacks 'the' . . . concept of squirrel. But there is no reason to think that having [that] belief requires that animals have that specific concept, nor that lacking the canonical concept of squirrel means that they lack any concept whatsoever" (Allen 1999, 35–36).

Thus a further question arises when considering the plausibility of the generality constraint in this context; namely, do we really apply the *same* exact mental concept to ourselves as we do to others? Many assume that the concept PAIN or VISUAL EXPERIENCE that one might attribute to oneself is the same as the concept that one attributes to another. It is then argued that since many animals cannot attribute the latter, they cannot attribute the former. However, this reasoning is highly questionable. For one thing, if my account of concept possession is viable, then it may just be that an organism O has a *better* understanding of a mental concept C when self-attributing C than when attributing C to another. And so, contra Seager and the generality constraint, it would be perfectly reasonable for an animal to be able to conceptualize its own mental states *to some extent* without being

able to conceptualize, to the same extent or in the same way, another's mental state.[15]

Moreover, isn't there a difference between MY PAIN and YOUR PAIN, or MY VISUAL EXPERIENCE and YOUR VISUAL EXPERIENCE? That is, since we only directly experience our own conscious mental states, it might simply be that the concept MY always implicitly accompanies my mental-state concepts. This seems reasonable at least in the sense that the process of concept *acquisition* is presumably quite different in each case. The lion or chimp has more immediate first-person access to its own mental states and thus can acquire concepts of its own mental states in a more direct way. However, acquiring and then attributing mental states to others (such as through YOUR PAIN or YOUR BELIEF) involves an additional, or at least different, inferential process. To put it bluntly: it is normally *much harder* to know about another's mental states than it is to know about one's own mental states. There is a reason why the problem of *other* minds is an age-old problem in the history of philosophy. One need not hold some radical Cartesian infallibility view to appreciate this point, since even the most anti-Cartesian skeptic will typically acknowledge that the admittedly fallible access to our own minds (as opposed to other minds) is at least *more* immediate or privileged in some important sense.[16] For example, I need not interpret my *behavior* in any obvious way when I think that I have a desire for food or a pain in my back. Thus if there really are two distinct concepts involved in self-attribution and other-attribution of mental concepts (because of the "my" and "your" qualifiers), then it is not even clear that the generality constraint is violated when an animal cannot other-attribute a mental state that it can self-attribute. And this would be yet another reason not to accept the logic that failure of some animals to other-attribute mental concepts means that they cannot attribute them to themselves. Attributing mental states to others seems to involve additional cognitive abilities, such as making certain inferences based on behavioral evidence, which some animals may not have or may have more difficulty acquiring. But this does not mean that they are incapable of having mental concepts at all or self-attributing them.

It is also worth mentioning that many of those who resist attributing concepts to animals tend to hold the rather strong view that links concepts to "inferential role" in reasoning and thus in connecting mental states to each other. Those who hold other views, such as an informational or teleological conception of the mental, may more readily permit additional concept attribution. In relation to the HOT theory, my own view again is that inferential role, reasoning, and language use most clearly appear at the level

of *conscious* HOTs. This is the more sophisticated level at which reasoning and first-person reporting occur.

In any case, if I am right, this bodes well for HOT theory, which only requires an (unconscious) I-thought with a self-attributing mental concept for an animal to have conscious mental states. I should add that it may indeed be that having I-thoughts (or HOTs about one's own mind) requires having thoughts about outer *objects* or *bodies*, for somewhat Kantian reasons. But having thoughts about other *minds* or *mental states* is different.

Much of the foregoing argument once again raises the general question of under what conditions is it reasonable to attribute any concept to a nonlinguistic animal. But we have already put forth CONPOSS and have applied it to infant concept possession. We can now see how this strategy works for animals. At the least, being able to differentiate between and recognize objects or properties (to some degree) is enough for conceptual understanding of the outer world. Similarly, being able to differentiate between and recognize mental states (to some degree) is enough for possessing mental concepts. In addition, a related core idea in the animal cognition literature is that the attribution of concepts is justified if evidence supports the presence of a mental representation that is independent of solely perceptual information (Allen and Hauser 1991). Similarly, some use "behavioral versatility" or "stimulus independence" as good evidence in support of animal consciousness (Griffin and Speck 2004; Newen and Bartels 2007). If an animal adjusts its behavior appropriately in response to novel and unpredictable challenges, it seems more likely that it is consciously thinking about its situation than when it responds uniformly. Fixed and rigid responses to stimuli seem to indicate a lack of conceptual representation. When one has concepts, one is thus able to form thoughts that contain those concepts. Indeed, Allen has proposed the following account:

> An organism O may reasonably be attributed a concept of X (e.g., TREE) whenever:
>
> (i) O systematically discriminates some Xs from some non-Xs; and
> (ii) O is capable of detecting some of its own discrimination errors between Xs and non-Xs; and
> (iii) O is capable of learning to better discriminate Xs from non-Xs as a consequence of its capacity (ii). (Allen 1999, 36–37)

Of course, if we are going to use it as a guide for attributing *mental* concepts to animals, then the X in question cannot be TREE or some external object but must be a mental concept (of which an animal arguably has at least a partial conception). We have seen that many animals seem to be able to meet condition (i); that is, they are able to distinguish one mental state

from another, such as seeing from not-seeing at all (or hearing), or remembering and not-remembering. For some, this may even be enough for an animal to have a mental concept, since it shows at least some understanding of those concepts by way of comparison. Nonetheless one could also make a case that some of the previously mentioned experimental results indicate an ability to meet clauses (ii) and (iii). If so, then an even stronger case can be made for possession of mental concepts. For example, any time a chimp or dolphin or rat detects its own error in, say, a metaconfidence or metamemory task and then goes on to perform the task better on that basis, it seems to have met (ii) and (iii). So when an animal learns to improve its performance on memory or confidence tasks, it seems to have understood how to *better* discriminate its own mental states from one another. Allen's own example (1999, 38) of pigs' "backout behavior" seems suited for this purpose. He describes cases where pigs display a self-monitoring of performance. Some pigs would attempt to back away from the choice they had made after committing to a response they had given (on, say, a same/different perceptual task), but before any feedback was provided.

In any case, I also find little reason to accept thesis (2).

8.3 Lloyd Morgan's Canon and Parsimony

Much has been made recently about how considerations of "parsimony" and Lloyd Morgan's Canon impact mental state and concept attributions to animals. The oft-quoted Morgan's Canon says that "in no case may we interpret an action as the outcome of the exercise of a higher psychical faculty, if it can be interpreted as the outcome of the exercise of one which stands lower in the psychological scale" (Morgan 1894, 53). On the surface, the canon seems often to favor a less-sophisticated behavior-reading hypothesis rather than a mind-reading interpretation of the evidence.

However, many commentators have noted numerous problems with this conclusion, as well as significant ambiguity in the canon itself (Bekoff and Allen 1997; Sober 1998; Allen-Hermanson 2005; Montminy 2005; Fitzpatrick 2008). Overall it remains unclear how to interpret Morgan's Canon, how it should be used to settle the debate surrounding animal mind reading, and how it relates to the associated notions of "parsimony" or "simplicity."

I wish to focus on the following points:

(1) Like other authors, Montminy (2005) rightly points out that one difficulty with Morgan's Canon has specifically to do with how to interpret the terms "higher" and "lower" in the canon. Various interpretations seem either incorrect or useless. But taking a cue from Bennett (1991), Montminy argues that what really matters is

how *concepts* should be attributed to animals, thus distinguishing between thinkers and nonthinkers. Thus the main issue is what *justifies* concept attribution to animals. I have also tried to focus on this aspect of the issue, such as applying CONPOSS to the available evidence. As we have also discussed, Bennett and Montminy emphasize that crediting animals with concepts depends on flexibility of behavior given the same stimuli; for example, an animal may seek food in many different ways as opposed to the same rigid response regardless of condition. It is here that, for example, discrimination and flexible recognition are central. Thus a dog will over time act differently with respect to bones depending on what else it detects. This serves as evidence that dogs have BONE. Much the same goes for FOOD, DANGER, SHAPE, and so on.

(2) Allen-Hermanson (2005) also critically examines the "higher" and "lower" notions in Morgan's Canon. He offers three interpretations of Morgan's Canon: the Metarepresentational Canon, Sober's Canon, and the Supervenience Canon. He rejects the first two. Although he rejects the Metarepresentational Canon as the only true interpretation of Morgan's Canon, there do seem to be cases where attributing metarepresentations is the best interpretation. We have already seen this with respect to metamemory, uncertainty monitoring, and metaconfidence judgments. The Metarepresentational Canon says that Morgan's distinction between higher and lower faculties should be understood as the difference between higher- and lower-order faculties. Of course, Allen-Hermanson correctly holds that not every dispute in this area comes down to a choice between first-order and higher-order mentality. There are also sometimes choices between attributing first-order mentality and offering a purely behavioral or physicalist account. Thus Morgan's higher/lower distinction does not exclusively align with either the mental/physical or the higher-order-mental/lower-order-mental distinction. The basic general lesson, if any, seems to be "don't go high if you can go low" (Allen-Hermanson 2005, 615).[17]

In the end, however, Allen-Hermanson takes "higher" to mean something like "supervenient," "emergent," or "nonreductive." He bases this on careful analysis of Morgan's own texts. For all I know, he may be correct as to Morgan's intentions and texts. Nonetheless I find this view a bit odd. It seems unusual to suppose that what seems primarily to be the *epistemological* or *methodological* nature of Morgan's Canon should be so dependent on a *metaphysical* view of consciousness and the mind–body problem. Do we need to settle the mind–body problem *before* using at least some form Morgan's Canon to guide research in cognitive ethology? It would seem not, especially since we are already referring to the mental states of animals.

Yet Allen-Hermanson (2005, 625) explicitly alludes to Levine's explanatory gap. Nonetheless we saw in chapter 2 that the explanatory gap is primarily an epistemological problem, even according to Levine himself. Thus the analogy offered by Allen-Hermanson is unclear in the sense that he first attempts to explain Morgan's higher/lower language in terms of a metaphysical view but then later claims that the problem at hand is akin to the more epistemological explanatory gap.

(3) Perhaps the most thorough recent critique of Morgan's Canon can be found in Fitzpatrick 2008, which argues that we should abandon Morgan's Canon because it is not in fact a good methodological principle at all. Nonetheless I think that Fitzpatrick is correct in first offering a reasonable interpretation of the terms "higher" and "lower" as "more sophisticated" and "less sophisticated," respectively. Thus we can think of the canon as guiding us to adopt the less sophisticated of any two interpretations of animal behavior. However, the question still remains as to how to determine which of two interpretations is less sophisticated. If we construe Morgan's Canon as a kind of principle of parsimony (or simplicity), then we still have the problem of specifying simplicity *relative to something*. Should it be the number of *processes* attributed? The number of *objects* or entities? The required amount of *memory*? And so on. It remains unclear why one version of the canon should be adopted over another. We might endorse something like a *ceteris paribus* clause, but of course it is arguably *never* true that *everything* else is equal (Fitzpatrick 2008, 232–233).

Moreover, as Morgan himself recognized, in some cases the "simplest" explanation for an animal's behavior *is the most anthropomorphic one*. Now, even if we disagree with Fitzpatrick's strong negative conclusion, the important point here is simply that those who offer behavior-reading interpretations of animal behavior often do so at the expense of other (at least) equally plausible notions of simplicity (Povinelli and Vonk 2006; Carruthers 2008). And this is precisely what leads Fitzpatrick (2008) to conclude that we are better off simply using the best *evidence* available to *justify* mental-state attributions regardless of the level of sophistication. What is really doing the work is the evidence for, say, concept or thought attribution, and thus the canon itself cannot serve as part of the rationale or justification for attributing or withholding mental states. This is similar to the strategy recommended by Montminy. At worst, we should remain agnostic in some cases.

Let's go further on this point. Now it may indeed *always* be *possible* to construct creative and often highly elaborate alternative first-order or even purely behavioral explanations for any given set of animal behaviors (Povinelli and Vonk 2006; Carruthers 2008). But just because an alternative

explanation is possible, it doesn't follow that it is the best or simplest explanation. Fitzpatrick (2008) makes a similar point: "One can always come up with some deflationary explanation for any putatively intelligent behavior . . . including any suite of human behavior. . . . Such deflationary explanations are always possible, though they will often be extremely implausible" (239–240).

Browne (2004) also explains that Morgan's Canon is thus not quite the same as following a law of parsimony. Browne, who is no friend of HOT theory, rightly recognizes that "it is parsimonious to explain similar, complex, stimulus-response patterns by similar psychological mechanisms" (2004, 648). So when various animals perform in ways similar to humans on, say, metacognitive tasks, "it is *un*parsimonious to adopt one kind of lower-level explanation for the animal's response on one task and a different kind of lower-level explanation for the animal's response on [another] task" (643–644). Browne thus seems to have in mind what I consider to be a reasonable *analogical* or *explanatory* notion of simplicity; that is, we ought to attribute mental states to animals (and thus explain their behavior) when they behave similarly to humans under similar conditions. Tomasello and Call (2006, 380–383) also argue that their opponents must often propose numerous extremely complex alternative explanations and learning scenarios to account for the same data. Thus considerations of parsimony can actually point *toward* the mind-reading hypothesis in many cases, and behavior-reading accounts often become quite ad hoc.[18]

(4) Let me give one concrete additional example. To resist the mind-reading interpretation of various experimental results, Carruthers (2008, 67) finds it necessary to posit the existence of a "gatekeeping mechanism" (as well as several other capacities) to try to explain how it is possible to interpret the evidence cited earlier as involving only first-order (unconscious) mental states. The gatekeeping mechanism acts on competing goals to control behavior. But, as Lurz has pointed out to me in e-mail correspondence, it is first still unclear why the gatekeeping mechanism is not itself metacognitive, despite Carruthers's claim to the contrary. Second, even if one accepts the first-order explanation offered by Carruthers, it is not clear that it is simpler than a mind-reading explanation, at least in many important and relevant respects. Positing additional animal mechanisms and capacities seems to run *against* parsimony, especially when the animal behavior in question is very similar to human behavior under similar conditions. Along these lines, Allen-Hermanson (2005) notes that "it might [sometimes] be more appropriate to posit a higher degree of representational complexity, in return for fewer system states" (2005, 617). This seems an apt point in

response to Carruthers 2008. Indeed, Carruthers himself seems to concede that the human and animal behaviors "parallel" each other in many important ways.[19]

In any case, perhaps it is best to interpret Morgan's Canon loosely in terms of degrees of cognitive sophistication. I side with those who hold that attributing higher-order mental states (and consciousness, for that matter) is, at least very often, the more justified and parsimonious move to make. At a certain point, arguing that virtually any, often extremely elaborate, alternative explanation to higher-order psychological explanations is preferable stretches beyond recognition any reasonable considerations of simplicity and rationality. This is all the more persuasive when one considers various other similarities between humans and other animals, such as evolutionary history, behavior, and brain structure.[20]

(5) Finally, much of the foregoing argument allows us to respond more directly to Bermúdez's (2003) argument against the very possibility of nonlinguistic animals having any metacognitive capacities whatsoever. Although Bermúdez does allow for nonlinguistic first-order "instrumental thinking," he argues that metarepresentational thought (or "intentional ascent") requires a public language with at minimum a combinatorial syntax and semantics. He claims that intentional ascent requires a suitable vehicle that is held in mind, so that a higher-order thought can be directed at a first-order thought. According to Bermúdez, only language can provide such a vehicle, because only language has the requisite structure to do the job, namely, to allow the use of inferences in reasoning. As he puts it: intentional ascent requires semantic ascent.

A number of problems arise, however: First, like many others, I reject Bermúdez's view that concept possession is so closely tied to linguistic competence, as is clear from my criteria in CONPOSS. I have also urged that concept possession is a matter of degree. Thus, for example, when Bermúdez (1998) speaks approvingly of "bodily self-awareness," I think we should understand it as a genuine form of self-concept, albeit a primitive or minimal one, which can figure into HOTs. Similarly, an animal's or infant's concepts of CAUSATION or NEGATION are genuine concepts, not mere "proto-concepts." Moreover, to rule out any and all nonlinguistic conceptual thinking seems a bit strong and unmotivated.

Second, although his discussion does not mention HOT theory, there is a clear respect in which Bermúdez has in mind the more sophisticated and explicit "introspection" or "reflection" as paradigmatic metacognitive states. For example, he says that holding thoughts in mind means "entertaining them consciously and considering how they relate to each other logically

and evidentially" (159). Bermúdez may be right, as I have also noted, that this is the level at which much conscious inference and reasoning take place. However, as I have emphasized repeatedly, there is also a more modest form of metacognitive state, namely, an *unconscious* HOT that accompanies each conscious state. Animals need not have introspective capacity to have this kind of metathought; there can be nonlinguistic creatures that have *implicit* higher-order propositional attitudes. So even if introspection always implies linguistic ability (which I also question), unconscious HOTs do not. Moreover, this tells against the requirement imposed by Bermúdez that a first-order state must be "held in mind" by a creature so that a higher-order thought can be directed at it. First-order states are only held in mind during introspection, not during typical unconscious HOTs when having outer-directed conscious states. Thus nonlinguistic conceptual thoughts in the form of unconscious HOTs are still possible for nonlinguistic creatures.[21]

Third, I find it odd to suppose that first-order intentional states can have the appropriate structure for the various reasons given by Bermúdez (such as the ability to determine modes of presentation), but second-order states cannot. How is it possible to assign structured first-order propositional attitudes to an animal but never in principle a structured second-order thought? Put somewhat differently, if public language is required for higher-order propositional attitudes, then why isn't it also required for first-order inferences and reasoning about states of affairs (Lurz 2007)?

Fourth, as Lyyra (2005) points out, Bermúdez does overlook some empirical evidence supporting the notion that animals have metathoughts, such as the results from uncertainty monitoring discussed earlier. Lyyra also explains that if Bermúdez is correct, his view would seem to rule out metacognitive capacities in aphasics, patients who have seriously impaired ability to express propositions in speech or writing (cf. Lurz 2007). However, aphasics are able to pass the nonlinguistic version of a false-belief test (Varley 1998), which would even seem to indicate the presence of metathoughts about *others*. It would surely be odd to suppose that aphasics are not capable of having metarepresentational thoughts.

8.4 An Aside on Autism

Many of the points raised here and the previous chapter also apply to autistic humans, so this is a natural place to address the topic briefly. I think that autistics are capable of having I-thoughts and thus also have conscious mental states according to HOT theory and the WIV. Some theory-theorists, however, have claimed that individuals with autism are "mind-blind" in

more significant ways and are virtually incapable of mindreading and thus perhaps even metacognition or at least some form of self-consciousness (Baron-Cohen 1995; Carruthers 1996; Frith and Happé 1999). Given his parallel arguments regarding animals and infants, Carruthers seems explicitly committed to the equally startling and highly counterintuitive view that autistic children lack conscious states. A defender of theory-theory reasons that if autistic subjects lack a theory of mind and if such a theory is also required for self-awareness, then autistic individuals should be "as blind to their own mental states as they are to the mental states of others" (Carruthers 1996, 262), and "they lack phenomenally conscious mental states" (Carruthers 2000, 202). I have already responded at length to similar worries regarding infants in the previous chapter, especially with respect to the false-belief task. In this chapter, I have fended off similar charges with respect to animals, presenting evidence for I-thoughts in animals, including both self-concepts and mental concepts. Now I focus on autism, which has become a much-discussed psychopathology, especially in connection with mind reading and metacognitive deficits.

Autism is a developmental disorder that affects a child's ability to develop social skills and engage in social activities. It is sometimes thought of as a more serious version of a spectrum of cases called Asperger's syndrome. Asperger's proper lies at the mild end of the spectrum, but varying degrees of impairment exist. Researchers widely agree that autistic humans have a number of clear deficits, such as impaired empathizing skills and deception detection. Autistics also exhibit a pronounced lack of imagination and ability to pretend, as well as significant difficulties with false-belief and joint-attention tasks (Leekam 2005). Thus it seems clear that autistic humans do indeed have particular difficulty with mind reading (Baron-Cohen 1995; Frith and Hill 2003; Nichols and Stich 2003). Primary symptoms include abnormalities in social development, in communication development, and in pretend play. There is typically a lack of normal eye contact and gaze monitoring, along with a lack of normal social awareness and responsiveness, such as would normally occur when one is embarrassed or sympathetic to another's embarrassment (Hillier and Allinson 2002). Subjects often display repetitive motor mannerisms such as hand waving and rocking. So, for example, Baron-Cohen (1995) argues that various mechanisms are impaired in the mind of autistic humans, such as major impairment of what he calls the "Shared Attention Mechanism."[22]

My main conclusion here will be that there is little reason to suppose that autistics do not have any, or even very few, I-thoughts or metacognitive states. It is also worth revisiting the dubious thesis discussed in section

8.2.3, namely, that having I-thoughts requires having thoughts directed at others' mental states.

A number of points come to mind:

(1) One initial problem with the literature is that some authors who argue for a deficiency in "self-consciousness" among autistic individuals leave the term undefined. As we have seen, self-consciousness, self-concepts, I-thoughts, and concept possession can come in degrees. At the most sophisticated level, there is introspection or reflection. Even if there are deficiencies in introspection, it does not follow that there are no I-thoughts or metacognitive states at all. To say that self-consciousness is impaired in some ways is one thing, but it is quite another to hold that there is no self-consciousness at all. Even the term "mind-blind" is ambiguous. As we have seen, there are varying degrees of HOTs, ranging from unconscious HOTs to reflective conscious HOTs.

Frith and Happé (1999) are perhaps most guilty of this ambiguity. They often use the terms "self-consciousness" and "introspection" interchangeably. For example, they talk about "introspective awareness" (1) and use "impaired self-consciousness" and "introspective awareness" on the same page (8), and then "reflective consciousness" later (10). But even if an autistic subject lacks sophisticated introspective capacity, this does not rule out a more modest kind of self-awareness or self-consciousness. We must keep in mind the distinction between conscious HOTs and unconscious HOTs, not to mention to various degrees of self-concept. Thus when Frith and Happé (1999) ask in their paper's title "What is it like to be autistic?" we need to be clear about whether they mean "What would a mind without *introspective awareness* be like?" (8; italics mine) or something much stronger like "What would a mind without any kind of self-consciousness be like?" As we saw for both animals and infants, there are often good reasons to suppose that a conscious organism can have less-sophisticated I-thoughts without having introspective states.

Moreover, it is one thing to suppose that autistic humans have *abnormal* or *different* self-consciousness, but quite another to claim that there is *no* self-consciousness at all. Indeed, even Frith and Happé (1999, 11–14) quote numerous cases of first-person reports from autistics. These reports actually seem to *favor* the presence of a self-concept, including some uses of "I." Many of the cases come from Hurlburt, Happé, and Frith (1994) and seem to indicate that autistic subjects are indeed capable of reporting their current thinking and feelings (cf. Nichols and Stich 2003, 185–187). Surely this indicates the presence of some form of self-awareness or I-thoughts.

(2) There also seem to be numerous cases where autistic subjects engage in deep meditation and prolonged focusing of attention on inner feelings or images (Ridge 2001, 331–333; Frith and Happé 1999, 14–16). Thus we have examples where introspective ability is sometimes even *greater* than normal, not to mention the admittedly unusual case of Temple Grandin (1995), who is a professor with a Ph.D. in animal science. One might turn the tables on theory-theory and argue that instead of a lack of mind-reading skills negatively impacting one's metacognitive ability, such an intense self-awareness might cause subjects to lack the typical awareness of others. That is, the self-preoccupation of some autistic individuals might even explain their lack of mind-reading skills. Many of the main deficits in question, such as impaired empathizing skills, lack of imagination, and difficulties with joint attention, might result from a *heightened* sense of introspection. To use an analogy: psychologists often trace the lack of empathy in serial killers to greater-than-normal self-absorption and narcissism. I am not comparing autistic individuals to serial killers but merely pointing out, by analogy, that abnormalities in introspection and self-awareness can also profoundly affect some mind-reading abilities.

(3) Autistic individuals do poorly on false-belief tasks, as is emphasized by many authors including Baron-Cohen (1995) and Nichols and Stich (2003). But we saw in the previous chapter that this was not really a problem for infant consciousness or concept possession overall. For example, it is important to distinguish between belief states and volitional states, such as goals and desires. Autistic children have some ability for desire attribution despite problems with belief attribution. Baron-Cohen (1995, 63–64) acknowledges as much when he cites evidence indicating that what he calls the "Intentionality Detector" is intact in autistic children. They are able to identify desires and goals of others and understand that desires can cause emotions.

Much the same goes for the so-called appearance-reality task. Children are presented with an object that appears to be one thing (such as a rock) but really is something quite different (a sponge). Autistic children do more poorly than their normal counterparts. Once again, these results are used by theory-theorists to infer not only that the children do not understand the appearance-reality distinction but also that autistics lack awareness of their own mental states.

Again, it can often sound as if Carruthers and others have in mind *reflective* self-awareness, which is not required for HOTs or conscious states. Furthermore, as Zahavi (2005) points out, the foregoing line of argument confuses "necessary and sufficient conditions" in ways that we have

already seen with respect to the false-belief task. He explains: "Even if success on the appearance-reality task justifies [i.e., is sufficient for] ascribing self-awareness to the person in question, one cannot conclude that somebody who fails on the task will lack [all forms of] self-consciousness. Such a conclusion would be warranted only if the ability to distinguish appearance and reality were a necessary requirement for possessing self-awareness, and no argument has yet been put forth to show that that should be the case" (195).[23]

Although I would agree with Zahavi that Carruthers's HOT theory cum theory-theory rules out autistic consciousness, I disagree that HOT theory itself really does entail a lack of phenomenal consciousness on the part of autistic subjects. It is the result of an outright confusion between "reflection" and other weaker forms of self-awareness. In addition, one might even be tempted to agree with Zahavi (2005, 194) that any theory that really does have the consequence that autistics do not experience any taste, auditory, or bodily sensations should serve as a *reductio* of the theory. Autistics would literally have to be viewed as zombies incapable of any feelings, emotions, or sensations. Autistic subjects, to be sure, have some severe abnormalities, even not feeling pain in some cases, but this does not warrant the exceedingly strong conclusion that autistics have no phenomenally conscious states at all. Once again, impaired or deficient self-consciousness, even seriously so, is not the same as having *no* self-consciousness. Similar objections equally apply to Frith and Happé, as we saw earlier.

Thus the evidence once again shows that there is no reason to hold thesis (1) from section 8.2.3: Having I-thoughts requires having thoughts directed at others' mental states. This thesis is a central part of theory-theory, but it does not seem to be the case that the various mind-reading deficits mentioned earlier are *equally matched* by or cause corresponding metacognitive deficits.

(4) Additional evidence comes from the area of memory research. Although the relationship between autism and memory (and temporal awareness) is certainly fascinating and complex (Boucher 2001), some evidence shows that much working memory and long-term memory remain intact. It is not as if autistics are thought of as severe amnesiacs with little or no episodic memory. In one major study (Farrant, Boucher, and Blades 1999), subjects were given a test to remember a list of numbers and then asked what strategy they used to remember them, that is, how they went about memorizing the list. Children with autism performed well on the task, on a par with normal children. Thus metamemory is at least somewhat intact, and as we saw in connection to infants and animals, there would thus

seem to be an I-concept and mental concept present. For all the foregoing reasons, then, there seems little reason to withhold attributions of HOTs to the autistic. Thus I disagree with Frith and Happé (1999) that individuals with autism "lack the cognitive machinery to represent their thoughts and feelings *as* thoughts and feelings" (7).

To sum up: it is at the least premature to claim that most animals (and autistics) cannot have I-thoughts. At best, there seems to be growing evidence that most animals can mind-read and have I-thoughts to some extent, depending on their degree of conceptual sophistication. And even if mind-reading others is not found in many animals, it does not follow that no I-thoughts are present. In some cases, more research needs to be done, and we should perhaps be content to remain agnostic in some specific cases. Finally, this means that if HOT theory is true, we also need not deny that most animals (or autistics) are conscious. I conclude thus far that the Animals Thesis is not only true but consistent with the HOT Thesis.

8.5 Animals and Conceptualism: The Continuity Argument

The remaining task for this chapter is to show that the Animals and Conceptualism Theses are consistent. This requires rebutting an argument that explicitly relies on the premise that the perceptual contents of animals are nonconceptual. Recall the following definitions from chapter 6:

(CON) Whenever a subject S has a perceptual experience *e*, the content *c* (of *e*) is fully specifiable in terms of the concepts possessed by S.

(NC) Whenever a subject S has a perceptual experience *e*, the content *c* of *e* is at least partly specifiable in terms of concepts *not necessarily* possessed by S.

Some authors have argued that nonconceptual content is needed to explain the continuity between human and animal perception (Evans 1982; Dretske 1993; Hurley 2001; Peacocke 2001a,b). Surely animals and humans have something perceptually in common when, say, they each consciously perceive a brown tree. Yet animals lack the relevant concepts, and so the content of *our* perceptual representations must equally be nonconceptual. Peacocke puts it thus:

> Nonconceptual content has been recruited for many purposes. In my view the most fundamental reason . . . lies in the need to describe correctly the overlap between human perception and that of some of the nonlinguistic animals. While being reluctant to attribute concepts to the lower animals, many of us would also want to insist that the property of (say) representing a flat brown surface as being at a

certain distance from one can be common to the perceptions of humans and of lower animals. . . . If the lower animals do not have states with conceptual content, but some of their perceptual states have contents in common with human perceptions, it follows that some perceptual representational content is nonconceptual. (Peacocke 2001b, 613–614)

Following Speaks (2005, 382), we can put the argument as follows:

(1) Animals possess no concepts.
(2) The contents of the perceptions of animals are nonconceptual. From (1).
(3) Animals and human beings are related to the same kind of content in perception.
Therefore,
(C) The contents of human perceptions are nonconceptual. From (2) and (3).

Thus this so-called *continuity argument* is cited as a major motivation for holding NC and rejecting CON (Byrne 2005). In my view, premise (1) is false. Thus premise (2) and the conclusion are also false. Premise (3), however, could still be true if "the same kind of content" is taken to mean "the same kind of *conceptual* content." Let us first see how Brewer and McDowell handle this problem.

8.5.1 Brewer and McDowell on Animals

I find the responses by Brewer and McDowell to the continuity argument to be unsatisfactory. Brewer (1999, 177–179), for example, offers a very weak defense of conceptualism with regard to animals and, even worse in my view, seems open to the idea that animal (and infant) conscious perceptions have at least some level of nonconceptual content (or "non-conceptual perceptual sensitivity"). He asks rhetorically: "What kind of connection must the non-conceptualist make between non-conceptual perceptual content and conceptual thought; or what kind of connection must the conceptualist make between conceptual thought and non-conceptual perceptual sensitivity?" (179). I think Brewer's approach is mistaken because the conceptualist is then in virtually the same predicament that he accuses the nonconceptualist of being in. A conceptualist should be able to do better and show why CON is more plausible than NC. Brewer is much too willing to construe this problem as a stalemate or trade-off, whereas I have at least tried to show that infants and animals have the requisite concepts to support HOT theory and, by extension, conceptualism. Brewer does, however, make the important general point that we should view the

continuity argument as parallel to the problem of infant concept possession. I agree and thus argued at length in the previous chapter that not only do infants have concepts, but infant perceptual consciousness is fully conceptual from the very beginning (given some sort of core nativism). I also show how concept acquisition is possible and so avoid the problematic transition from any alleged nonconceptual content to conceptual content. It seems reasonable to suppose that many of the same arguments in defense of CON apply equally to animals. Brewer is still correct, however, that the continuity argument cannot serve *on its own* as a motivation for introducing nonconceptual perceptual content.

McDowell's discussion of animals in *Mind and World* is also unsatisfactory and often unclear (1994, 63–65, 69–70, 114–123, 182–194). First, he concedes far too much regarding the nature of concept possession. He seems to agree with premise (1) of the continuity argument and unwisely follows Evans by endorsing a demanding notion of concept possession linked to linguistic use.

Second, partly due to his notion of concept possession, McDowell struggles to make any sense of animal consciousness and how it differs from human consciousness. According to him, animals do not have Kantian "spontaneity," and thus they do not have self-consciousness or the ability to reason. But since concept application is necessarily intertwined with perception, it becomes unclear just how McDowell can allow for animal consciousness at all, even pains and fears. He tells us that animals cannot have "objective experience," that is, experience of an outer world of objects. Other remarks only make matters worse, such as "we can say that we have what mere animals have, perceptual sensitivity to features of our environment, but we have it in a special form . . . taken up into the ambit of the faculty of spontaneity, which is what distinguishes us from them" (1994, 64; cf. 114).

But McDowell is often ambiguous as to what he means by "perceptual sensitivity to features of our environment," which we still allegedly "share" with animals. He obviously cannot mean conscious perceptions with nonconceptual content, because in his view, *our* perceptual content is thoroughly conceptual. But he also cannot mean perceptions with conceptual content, since he thinks that animals do not have concepts. Thus it is unclear whether or not premise (3) is false according to McDowell. On the one hand, he tries but fails to make sense of similarities between human and animal perceptions. On the other hand, he also makes clear that there is an important difference between humans and animals in terms of possessing concepts. At the least, it remains unclear how premise (3) could be true or

false on McDowell's account. Animals either have conscious perceptions of outer objects or do not—so which is it? McDowell noticeably avoids using the terms "conscious" or "experience" when describing animal perception, but he insists that there is "no Cartesian automatism in [his] picture" (116), and he has "no wish to play down the respects in which [animal] lives are like ours" (183). In addition, he seems to have missed the important Kantian point that coherent conscious perception *requires* outer perception. McDowell seems to think that this Kantian point applies only to humans.

Moreover, it is equally puzzling how animals can even have conscious pains, feelings, or any kind of "sentience," according to McDowell (70, 119–122). He uses obscure references to "animal life" and tells us that "animals are natural beings and no more" (70). Are we therefore *super*natural beings because we have concepts? Again, the lack of clear reference to consciousness is striking. McDowell invites similar criticism when he says, for example, that "feelings of pain or fear need not amount to awareness of an inner world" (119), and that animals lack PAIN. I find much of his discussion along these lines to be exceedingly cryptic. There are frequent metaphorical references regarding how the sentience of animals "actualizes itself" and is a kind of "proto-subjectivity" (as opposed to "full-fledged subjectivity"), not to mention an unhelpful digression into the views of Gadamer (McDowell 1994, 115). In contrast and with the help of HOT theory, I have tried to show how one can be a conceptualist with Kantian influences while also clearly endorsing animal consciousness. We are much better off unambiguously acknowledging consciousness for most animals and then showing that what mainly differentiates us from them is the degree of conceptual sophistication.

McDowell also seems in effect to be claiming that having conscious mental states involves having something more like introspection, which, once again, is not the case. He speaks of animals' lacking "self-critical thinking" (1994, 69), a "contemplative attitude" (117), as well as the ability to make decisions, have freedom, voluntarily guide behavior, and possess agency, reflection, and rationality. Some of these abilities are arguably also necessary for making ethical judgments. It may very well be that animals or most animals lack these abilities, but it also seems to me that they require conscious HOTs. But, as we have seen often, unconscious HOTs (or judgments) are enough for first-order conscious states and for concept possession.

Third, McDowell's use of language is telling. He regularly speaks of "mere" and "dumb" animals (1994, 69, 182), which recalls the language of Kant and other philosophers centuries ago but is certainly not prevalent today. It is worth mentioning that Leibniz also struggled with the issue of

"brute apperception" and "brute rationality." In my view, a good case can be made that Leibniz held both that humans are psychologically superior to animals in many important ways and that animals are both conscious and self-conscious (Gennaro 1999). Indeed, I think that Leibniz anticipated some of the issues raised by Kant, McDowell, and the HOT theory. He was certainly ahead of his time in endorsing unconscious mental states. In any case, it is puzzling why McDowell, who is often concerned about retaining animal–human continuity, especially in light of evolutionary development, would come so dangerously close to denying animal consciousness. Much the same goes for the continuity of an infant's development into adulthood.

McDowell's more recent attempts to clarify his position on animal consciousness in response to similar worries are equally unhelpful (McDowell 2002, 283, 299; 2008, 220–222, 234–237). In one place, he briefly acknowledges that there needs to be a middle-ground way of thinking about nonhuman animals, but he does little to develop such an account (2002, 283, 299). One could view much of this book as a sustained, detailed attempt to find such a middle ground against the background of HOT theory. In other places, McDowell reiterates that although infants and animals have "sensibility," they cannot engage in thinking (2008, 227). Once again, it is then unclear just how to separate out what infants and animals *share* with adult humans if they can have *no* concepts or thoughts at all.

8.5.2 Other Conceptualist Replies

So how *should* a conceptualist reply? The most obvious immediate answer to the continuity argument is that both premises (1) and (3) are highly questionable. First, regarding premise (3), at least *some* animal perceptions are presumably quite *unlike* ours, such as consciously perceiving certain smells. Nagel's famous bat comes to mind in this context. Thus a conceptualist should reject premise (3) if it is taken to imply that human and animal perceptions are similar in content across the board. Note, however, that rejecting premise (3) alone need not commit one to NC. The difference in perceptual content between humans and animals could simply be explained by differences in conceptual content. It could be that the concepts that animals lack are precisely those deployed in exclusively human perceptual experience. Or it could be that our concepts are much more fine grained than theirs, and this explains the perceptual differences in question. On the other hand, it is likely that some animals have more fine-grained perceptions than we do in some cases, such as the more fine-grained and perceptual sensitivity of many dogs with respect to smells. Either way, the

conceptualist should hold that when (or if) an animal really does have the same kind of conscious perception as a human, *both* contents are fully conceptual. In this sense, then, a conceptualist could agree with premise (3) but insist that the "same kind of content" is conceptual in nature.

A conceptualist should also reject premise (1). It is much too strong, and one need not be a conceptualist to think so. As we have seen throughout this chapter, there seems to be ample evidence for attributing concepts to animals. As a matter of fact, animals have more sophisticated concepts than we might otherwise have thought, such as mental concepts. It is therefore reasonable to think that most animals have many of the primitive concepts possessed by infants, such as OBJECT, TIME, SPACE, SHAPE, SIZE, NUMBER, and so on, not to mention quite a number of other basic concepts, such as FOOD and ANIMAL. Parallel experiments on animals strongly suggest a parallel with infants for these kinds of concepts. In addition to concepts like PREDATOR, many animals also seem to have at least some domain-specific understanding of concepts related to artifacts and tools, such as RIGIDITY and FUNCTION (Hauser and Santos 2007). For example, many primates, such as lemurs and rhesus macaques, recognize that shape, size, material, and orientation are relevant features of tool function, whereas color is not. Moreover, using the preferential looking-time method, for example, Hauser and colleagues found that different primate species keep track of individual objects placed behind a barrier, indicating an understanding of numerical concepts (Hauser, MacNeilage, and Ware 1996). Finally, as we have seen, even some who question the very existence of animal consciousness allow for animal concepts extremely far down the evolutionary scale (Carruthers 2005, chap. 12; 2009a). And obviously if premise (1) is false, then so would be the natural strong reading of premise (2), namely, that the contents of the perceptions of animals are *wholly* nonconceptual.

Premises (1) and (3) also appear to conflict. It does not seem that the nonconceptualist can have it both ways; that is, animals are so *different* from us with respect to concept possession, but so *similar* with regard to perceptual content. Byrne (2005) also notices this tension in the foregoing argument. He explains how odd it is, on the one hand, to suppose that animals are importantly like us perceptually but, on the other hand, to suppose that they are so radically unlike us cognitively that they cannot think, believe, or know anything (which surely requires conceptual content).

Speaks (2005) also rightly points out that slight variations of premises (1) and (3) could render them compatible if, for example, premise (1) were weakened to read, "Animals possess *some* concepts." This would support the inference to a weaker premise (2) that says that "the contents of *some of*

the perceptions of animals are nonconceptual." But this is only *compatible with* premise (3), because even if one allows for nonconceptual perceptual content in animals, one could then treat premise (3) as saying, "Animals and human beings are *often* related to the same kind of content in perception." In short, depending on how broad the scope of premises (1) through (3) is, questions arise as to the very validity of the argument.

Noë (2004, 184–189) also replies in ways consistent with the arguments of this and the previous chapters. For example, he recognizes the need to view concept possession as a matter of degree and holds that concepts can often be deployed implicitly in perception. He argues that we would not even credit a person or animal with the visual experience of an anteater if we did not believe the person or animal had ANTEATER. Perceptual experiences are, after all, paradigmatically intentional states. "It is difficult to understand how one could have an experience with a given intentional content without being in a position to *understand* that content" (189; italics mine). It is precisely this understanding that requires concept application or "seeing-as."

Thus what really differentiates us from most animals is not consciousness but the degree of conceptual sophistication we bring to experience, in addition to a host of other abilities such as reflection and reason. CON is true both for animals and for humans. Partly motivated by HOT theory, we should suppose that CON applies to any conscious organism capable of having conscious perceptions. It is important to keep in mind, of course, that there may well be some simple animals or insects incapable of having HOTs. In those cases, they would also not be conscious. Nonetheless, contra Carruthers, I think that the line (to the extent there is a clear line at all) is fairly low on the evolutionary scale.

At the least, there are many plausible replies to the continuity argument available to the conceptualist. CON is not refuted, and we have good reason to suppose that conceptualism applies equally to all conscious animals.

Overall, then, I conclude that the Animals Thesis is true and that it is consistent with the HOT and Conceptualism Theses. Moreover, the case for the Conceptualism Thesis is complete because it is consistent with the Infants, Acquisition, and Animals Theses. We are virtually finished with solving the Consciousness Paradox. There is, however, one remaining thesis to defend. I now turn to that task.

9 Into the Brain

In this final chapter, I defend the HOT-Brain Thesis, which says that there is a plausible account of how my version of HOT theory might be realized in the brain and can lead to an informative neurophysiological research agenda. Alternatively, HOT theory is related to, and consistent with, a number of leading empirical theories of consciousness. This involves delving further into the question of how my theory of consciousness might be realized in the brain. As I noted in chapter 1, I disagree with Revonsuo's claim "these theories [= HOT theories] have not had any major impact on the empirical study of consciousness" (2010, 189). In section 9.1, I present some basics on brain structure and function. I also discuss the problem of finding the so-called neural correlates of consciousness (NCCs). In section 9.2, I frame the main issue in terms of the question "How global is HOT theory?" That is, how widely distributed in the brain are conscious states? I argue that HOTs need not occur in the prefrontal cortex although HOT theory demands that conscious states be distributed to some degree. In section 9.3, I revisit the mereological issue explored in chapter 4 with much more specific emphasis on the WIV, the brain, and feedback loops. In section 9.4, I end with a discussion of the importantly related binding problem and the unity of consciousness. I argue that HOT theory and the WIV can accommodate various attempted solutions to the binding problem and can shed some light on the matter. I conclude with a somewhat speculative proposal regarding the overlap between the binding problem, the search for NCCs, and the hard problem.

9.1 The Neural Correlates of Consciousness (NCCs)

The search for the neural correlates of consciousness (NCCs) has become a major preoccupation among philosophers and scientists alike (Metzinger 2000; Blackmore 2004, chap. 16; Hohwy 2007). It has to do with

determining the exact relationship between brain activity and conscious experience. Narrowing down the precise brain property or properties responsible for consciousness is a far more difficult enterprise than merely holding the more generic belief in some form of materialism, as reasonable as that might be. For example, it is not even always clear just what kind of brain property could be responsible for consciousness. Is it a neural property, such as firing rates? Is there an important chemical component, such as a certain neurotransmitter? Are conscious states mostly locally represented in the brain, or are they more widely distributed? Is there a different NCC for state consciousness and creature consciousness? Before going too far, however, let us review a few basics of brain science.

9.1.1 The Brain: Some Basics

Most readers will be familiar with the general structure of the brain and have some knowledge of neurophysiology. But for those not familiar with brain science, a brief description of the parts of the brain and how neurons work is in order.[1]

The brain is divided into the left and right hemispheres, which are connected by an extensive band of nerve fibers collectively called the corpus callosum. The main brain structures of the neocortex include four lobes: the frontal lobe, the parietal lobe (top of the brain), the occipital lobe (in the back of the head), and temporal lobes (on the sides of the brain).

The cerebral cortex involves several major structures, such as the hindbrain, which includes the cerebellum. The cerebellum controls balance and some motor coordination along with the pons. The midbrain includes the reticular formation and the superior and inferior colliculus. The forebrain encompasses the diencephalon (with the thalamus and hypothalamus), the telencephalon (e.g., the basal ganglia, which include the amygdala; and the limbic system, which includes the cingulate gyrus and the hippocampus). The neocortex with its four lobes is also part of the forebrain (see fig. 9.1).

Functionally specific areas are well known to be essential for various mental abilities, such as the visual cortex, the auditory cortex, and various deeper structures such as the cingulate gyrus, the basal ganglia, the hippocampus, and the thalamus. The thalamus is a subcortical structure that sends and receives signals from the cortical areas, including the primary sensory areas responsible for vision, hearing, and feeling. This interconnected set of systems is sometimes called the thalamocortical system.

The visual cortex, for example, is responsible for vision and is located in the occipital lobe. The classic area is labeled V1, but other areas include V2

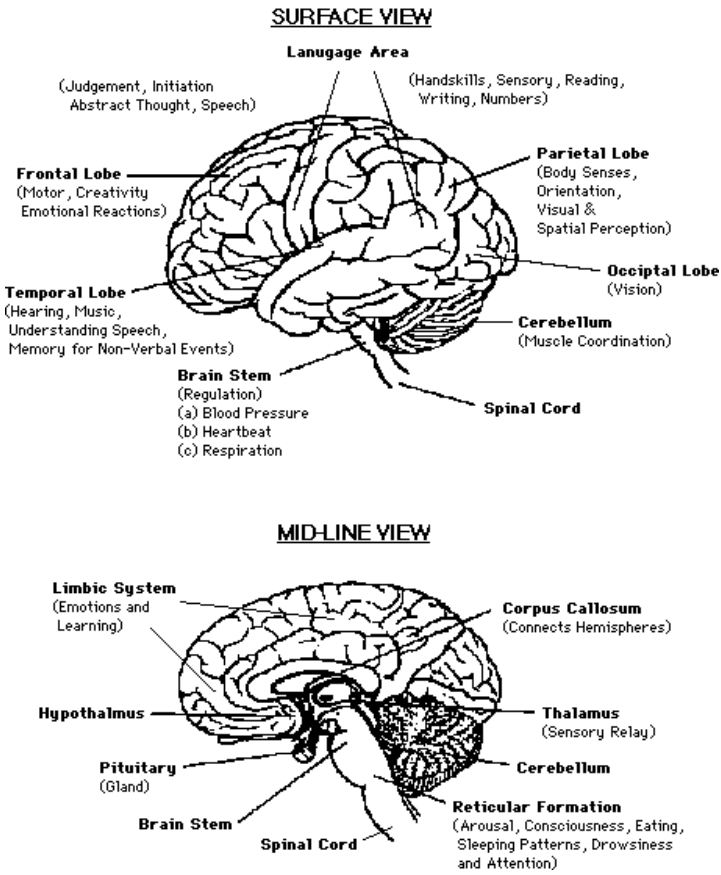


Figure 9.1

A map of major brain structures. The first image shows outer brain areas. The other is a lateral brain section showing additional inner structures. Reprinted with the express permission of TPN Inc. It may not be reused or reproduced without additional permission from TPN Inc. (http://www.tbi.org).

through V5, though V5 is also sometimes labeled MT. V5/MT, for example, is well understood to be responsible for motion perception. Other major brain areas include the motor cortex and the somatosensory cortex.

There are approximately 100 billion nerve cells, or neurons, in an average adult human brain. Neurons come in a variety of shapes, but they all have treelike projections called dendrites that receive synaptic connections (a synapse is the small distance between neurons). Dendrites project from the single longer projection, called an axon. The branchlike patterns of dendrites vary widely from neuron to neuron. In addition, given the incredible number of connections between neurons, there are many more neural connections than the mere number of neurons.

Neurons fire and communicate with one other via electrochemical activity. More specifically, neurons have a resting potential of –70mV, which is the normal voltage across the nerve cell membrane. If a neuron is excited by a neurotransmitter, a chemical released from a presynaptic neuron, then it causes a depolarization of the postsynaptic neuron. The depolarization causes brief changes in the neuron's permeability to potassium and sodium ions that, in turn, causes an electrical impulse (called an action potential) to occur at –50mV. The nerve cell fires at this point and not until this point. The firing of neurons is an all-or-nothing matter; that is, neurons do not fire to a lesser degree at, say, –60mV or to a greater degree at –40mV. The firing will then cause the release of neurotransmitters into the synapse of the postsynaptic cells, and then the cycle continues. Some synapses receive inhibitory signals that slow or stop activation of the postsynaptic neuron. Other synapses receive excitatory signals that increase the firing rate of the receiving neuron. All of this occurs over periods of tens to hundreds of milliseconds.

It is also important to recall that there are numerous feedback loops in the brain, also referred to as *recurrent processing* or *reentrant feedback*. That is, numerous neurons are connecting and transmitting not only from early processing areas to higher areas but also back from the higher areas to the early areas. A good example of this process can be found in the thalamocortical system, which is a dense network of reentrant connectivity between the thalamus and the cortex. This notion of reentrant connectivity has already played a central role with respect to defending HOT theory, but in particular to how the WIV might be understood in terms of the brain. Either way, however, due to feedback loops, many are inclined to think that the NCC must be somewhat distributed in the brain (Lamme and Roelfsema 2000; Edelman and Tononi 2000a, 2000b; Pascual-Leone and Walsh 2001), though just how widely is still quite controversial.

It is worth briefly noting some of the methods used to detect and measure brain activity in relation to various mental tasks. Positron-emission tomography (PET) is a way to construct brain images from the distribution of radioactivity following administration of a radioactive substance. PET scans measure brain metabolism and blood flow directly by measuring the atoms that emit positrons that are incorporated into oxygen or glucose molecules. Nuclear magnetic resonance imaging (MRI) measures the radio signals emitted by some atomic nuclei. The radiation emitted provides detailed information about the chemical nature of the nuclei. Functional MRI (fMRI) is a newer and more advanced method that allows for such imaging while the subject is engaged in various tasks.[2]

This is not merely a theoretical exercise. For individuals thought to be in a persistent vegetative state or under anesthesia, it is exceedingly important to be able to ascertain from a scientific, third-person point of view to what extent (if any) consciousness is correlated with specific NCCs (Alkire, Hudetz, and Tononi 2008; Revonsuo 2010, chap. 8). Errors in accurately determining when a patient is having conscious states, such as conscious pains, can have catastrophic results. Imagine suffering in excruciating pain while unable to move one's muscles in order to inform others. This is similar to "locked-in syndrome," which is a medical condition where brain damage has affected only motor functions and left the patient immobile and unresponsive to stimuli, but consciousness remains normal. Mashour and LaRock (2008) refer to this as the "inverse zombie problem," that is, cases of internally experienced consciousness without any behavioral sign, as opposed to the philosopher's "zombie," who is hypothetically not conscious but behaves in a manner indistinguishable from a conscious human (see chapter 2).

9.1.2 NCCs: Some Leading Candidates

One frequently cited candidate for the NCC, often called the "temporal synchrony" account, is offered by Crick and Koch (1990; see also Crick 1994; Koch 2004). The basic idea is that mental states become conscious when large numbers of neurons fire in synchrony with one another, say, oscillations within the 35 to 75 hertz range or 35 to 75 cycles per second. One oscillation per second is denoted as 1 hertz (Hz), but neural response times are often measured in thousandths of a second (or milliseconds, usually abbreviated as msec). A detailed survey of other contenders would be impossible to give here, but a number of other candidates for the NCC have emerged over the past two decades (Metzinger 2000), including reentrant cortical feedback loops in thalamocortical systems (Edelman 1989;

Edelman and Tononi 2000b), NMDA-mediated transient neural assemblies (Flohr 1995, 2000), emotive somatosensory hemostatic processes in the frontal lobe (Damasio 1999), and activation in the parietal cortex (Hardcastle 1995).

I will return to Edelman and Tononi later, but to elaborate briefly on Flohr's theory, the idea is that anesthetics destroy conscious mental activity because they interfere with the functioning of N-methyl-D-aspartate (NMDA) receptors. According to Flohr, then, the activation of the NMDA system is necessary for the mechanisms underlying consciousness. Flohr explicitly relates his theory to HOR accounts of consciousness by arguing that the NMDA synapse implements the binding mechanism that the brain uses to produce widely distributed representations to which HORs belong (Flohr 2000, 252–253). In addition to Flohr's own work, this connection has been noticed and discussed by others (Blackmore 2004, 230; Kriegel 2007b, 909).

Many others have also emphasized the importance of the role of neurochemistry in having various kinds of conscious states. Indeed, an entire anthology on the neurochemistry of consciousness has examined this often-ignored area of research among philosophers (Perry, Ashton, and Young 2002). Over fifty neurotransmitters have been discovered thus far. For example, acetylcholine seems to play a major role in the difference between sleep (and dreaming) and waking forms of consciousness. Dopamine seems to contribute importantly to attention and working memory. It is also widely acknowledged that some mental disorders, such as depression, dementia, and schizophrenia, result from abnormalities in the levels of neurotransmitters. Despite this fascinating line of thought, I will also not say much more about the neurochemistry involved and will mostly focus on the neural structures in question.

9.1.3 Three Clarifications

In any discussion of NCCs, we must avoid several problems and potential pitfalls:

(1) One issue is determining exactly how the NCC is related to consciousness. For example, although a case can be made that many of them are *necessary* for conscious mentality, it is unclear that they are *sufficient*. For one thing, many of the above candidates for NCCs seem to occur unconsciously, as well. Second, there are obviously other necessary background conditions that need to obtain for a given NCC to suffice for consciousness. Even pinning down a narrow-enough necessary condition is not as easy as it might seem.

A related worry has to do with the very use of the term "correlate." As any philosopher, scientist, and even undergraduate student should know, saying that "A is correlated with B" is rather weak (though it can be an important first step), especially if one wishes to establish the stronger *identity* claim between consciousness and neural activity. Even if a solid correlation can be established, we cannot automatically conclude that there is an identity relation. And many view the search for NCCs as somewhat neutral with respect to the metaphysics of mind. Perhaps A causes B or B causes A, and that's why we find the correlation. Most dualists could even accept such a view. Maybe there is even some *other* neural process C that causes both A and B. "Correlation" is not even the same as "cause," let alone enough to establish identity. Finally, some NCCs are put forth as candidates not for all conscious states but only for certain specific kinds of consciousness such as visual awareness.

Crick and Koch are faced with this problem, that is, whether temporal synchrony is *sufficient* for consciousness, merely *necessary* for it, or both. Ultimately they concede that the current data do not really support the conclusion that synchronization of neural assemblies constitutes a *sufficient* condition for production of conscious awareness. So this leads naturally to the question: what *else* is necessary that, perhaps along with temporal synchrony, might be sufficient for conscious experience?

(2) Recognizing the need for conceptual clarity, Chalmers (2000) does us the service of clarifying just how to understand a NCC. He presents several useful distinctions and offers a number of clear definitions (cf. Block 2007; Hohwy 2007). For one thing, as we have seen, we should distinguish between a mental state (or vehicle) and its content. Thus Chalmers arrives at the following definitions:

"A *content* NCC is a neural representational system N such that the content of N directly correlates with the content of consciousness" (Chalmers 2000, 20; italics mine).

"A *state* N1 of system B is a neural correlate of phenomenal property P if N's being in N1 directly correlates with the subject having P" (22; italics mine).

It is then important to recognize that any interesting NCC would at least need to isolate the *minimal* area in the brain responsible for a conscious state. Thus one finds the following:

"An NCC is a *minimal neural system* N such that there is a mapping from states of N to states of consciousness, where a given state of N is

sufficient, under conditions C, for the corresponding state of consciousness" (Chalmers 2000, 31; italics mine).

Others make similar remarks, such as when Block (2007, 489) explains that a "minimal neural basis is a necessary part of a neural sufficient condition for conscious experience," and when Koch (2004, 16) tells us that the NCC is "the minimal set of neuronal events and mechanisms jointly sufficient for a specific conscious percept."

The main point is to find a neural correlation that is a reasonably interesting subset of the entire brain activity at a given time (Chalmers 2000, 24–25; Block 2007). It would be much less informative, and perhaps trivial, to learn that the *entire* brain is sufficient for having a conscious state. In a similar vein, one might distinguish between the *core* and *total* NCC. The *core* neural basis of a conscious state is the part of the total neural basis that distinguishes conscious states from states with other conscious contents (cf. Zeki 2001). The *total* neural basis of a conscious state is itself sufficient for the instantiation of that conscious state (Block 2007, 482). We thus also need to distinguish the NCC from what might be called "enabling conditions," which refer to other aspects of a functioning body, such as proper blood flow and functioning lungs and heart (see also Block 2007, 485–486, for some discussion). There may be problems here as well, and perhaps we need new experimental approaches (Hohwy 2009). It is also crucial to design experiments with controls such that the only difference between a pair of trials is the presence of consciousness. We can then use fMRI to ascertain any neural difference between such cases.

(3) Last, let us also recall the state/creature consciousness distinction. One problem with Flohr's account, for example, might just be that he is really only focused on overall creature or organism consciousness in the sense of a creature being awake or aware of its surroundings. In some ways, this makes sense when thought of from the point of view of anesthesia and neurochemistry. The emphasis is on whether or not the *patient* is unconscious or when the *person* loses consciousness. Similar emphasis is found with respect to vegetative and coma states. For example, the literature widely notes that the reticular formation is the only localized brain area where lesions to it result in a complete loss of consciousness.

9.2 How Global Is HOT Theory?

Keeping the above neurological background in mind, a crucial issue is thus how one's theory of consciousness might be neurally realized in the brain in the sense of the "core state." What is the NCC according to HOT theory?

Although my theory is reductionistic in mentalistic terms, it is interesting to examine just how HOT theory (or my WIV) might be reduced further to neural properties. Recall that this might be viewed as a second-step reduction for a HOT theorist (chap. 2). Given the iterative structure of HOT theory and the WIV, one might reasonably suppose that the neural realization in question is fairly widely distributed in the brain, involving various areas of the brain.

Thus we might take the key question to be: how widely distributed, or "global," are conscious states? We saw in chapter 4 how some empirical evidence from the neurosciences can be used in support of the WIV. In this subsection, I argue that HOTs (or METs) need not occur in the prefrontal cortex, and thus, although HOT theory demands that conscious states are distributed to some degree, a more moderate global view is preferable, especially with respect to first-order conscious states. This remains the case even if one prefers to treat the lower-order state and the unconscious HOT as parts of a single unified state, as the WIV suggests.

It will first be useful to critique the views of Kriegel (2007b; 2009a, chap. 7) and Block (2007) and contrast them to my own. Although some of what follows is somewhat controversial, I think the current evidence supports my view.

Recall that I have already rejected Kriegel's self-representationalism (chap. 5), not to mention the stronger PSR. However, he argues that there is neuropsychological evidence for what he calls "cross-order integration" (COI) theory, which is another name for his "same-order monitoring theory" or "self-representationalism" (Kriegel 2007b, 2009a). Kriegel first explains that three elements are needed for NCCs in the COI theory:

(1) There is a floor-level (or first-order) representation.
(2) There is a higher-order representation of (1).
(3) There is the "functional integration" of (1) and (2) into a single unified state via some binding mechanism, perhaps along the lines of the temporal synchrony account.

Recall that, in Kriegel's view, the higher-order representations *are themselves conscious*, unlike in HOT theory. He also explains that the likely NCCs for the floor-level representations will depend on the modality, such as V1 to V5/MT for perceiving a moving patch of blue color. According to his view, this has to do with the *contents* of conscious states, as opposed to consciousness *as such*. Most crucially, Kriegel thinks that the likely NCCs for the second-level representations are in the prefrontal cortex (PFC).

Now one immediate problem with Kriegel's account is that his discussion of the "second element" reflects some sophisticated abilities, such as

executive functions and attentional control, which are better understood as *introspective* capacities as opposed to other, less-sophisticated forms of metacognition. Thus it might well be that *these* capacities are indeed subserved by PFC activity. However, there is little reason to think that they are required for any (or most) first-order conscious states. Thus, even if COI theory is committed to all metacognition involving PFC activity (because the higher-order representation is itself conscious when having a first-order conscious state), the same is not the case for HOT theory, which only requires an unconscious HOT to accompany a first-order conscious state.

It seems to me that this is an advantage for HOT theory with regard to the oft-cited problem of animal and infant consciousness. If COI theory *requires* PFC activity for *all* conscious states and HOT theory does not, then HOT theory is in a better position to account for animal and infant consciousness, since it is doubtful that infants and most animals have the requisite PFC activity (though perhaps some do). Kriegel's analysis is thus arguably not even an account of the NCCs of *first-order* conscious states at all. This may not, strictly speaking, falsify COI, since Kriegel could allow for the possibility of other NCCs for the higher-order representations (e.g., he mentions the anterior cingulate cortex as another possibility). I will return to additional evidence later.

In a provocative *Behavioral and Brain Sciences* target article, Ned Block (2007) addresses some related issues, but I think he is equally mistaken. Among other things, Block claims (a) that the "same order [= COI] view fits both science and common sense better than the higher-order [= HOT] view" (485), and (b) that "since frontal areas are likely to govern higher-order thought, low frontal activity in newborns [and presumably most animals] may well indicate a lack of higher-order thoughts about genuine sensory experiences" (485).

I disagree with Block on both counts. First, I have already argued at length in the previous two chapters that infants and most animals have the requisite psychological and conceptual abilities to have at least some HOTs. As we have also seen, COI theory really makes it *more* (not less) difficult to account for infant and animal consciousness. The HOR for Kriegel is more sophisticated (since it is itself conscious) than unconscious HOTs. At the least, Block is not justified in dismissing HOT theory without also rejecting COI. As Rosenthal (2007) explains, "According to standard same-order theories . . . that awareness [of the experience] is every bit as cognitive as on the higher-order thought hypothesis" (523). Moreover, given my extended critique of Kriegel's view in chapter 5, I certainly do not think that his view comports better with common sense. Second, although I agree with Block

that PFC activity is not necessary for having first-order conscious states, I disagree with the claim that "frontal areas are likely to govern higher-order thought" unless he primarily means *introspection* or conscious HOTs.

Now one might ask: what evidence is there of conscious states without PFC activity? There seems to be quite a bit. Here is a partial list:

(1) In a summary review, Tong shows that even though V1 may only be necessary for conscious visual experience, interaction between V1 and other areas (V2–V4 and MT) seems sufficient (Tong 2003; cf. Baars and Gage 2010, chap. 6).
(2) Basic conscious experience is not significantly decreased even when there is extensive bilateral PFC damage or lobotomies (Pollen 2008).
(3) When Transcranial Magnetic Stimulation (TMS) is applied to V5/MT, it causes subjects to experience moving phosphenes (Cowey and Walsh 2000; Pascual-Leone and Walsh 2001). TMS delivers an electromagnetic jolt to brain areas when placed on the scalp. The result is disrupted signals and created signals.
(4) Similar results are found for other sensory modalities, for example, in auditory perception, as discussed in Baars and Gage 2010, chap. 7. Although areas outside the auditory cortex are sometimes mentioned, there is virtually no mention of the PFC. We also find similar results for tactile awareness in the somatosensory cortex (Gallace and Spence 2010).
(5) Rafael Malach and colleagues show that when subjects are engaged in a perceptual task or absorbed in watching a movie, there is widespread neural activation but little PFC activity (Grill-Spector and Malach 2004; Hasson, Nir, Levy, Fuhrmann, and Malach 2004; Goldberg, Harel, and Malach 2006). Although some other studies do show PFC activation, this is mainly because of the need for subjects to *report* their experiences. The PFC is likely to be activated when there is *introspection*, not merely when there are outer-directed conscious states (especially during demanding perceptual tasks). A similar point is made by Crick and Koch (1995), who explain that visual representations must be sent to the frontal cortex to be *reported* and for subjects to *reason* about them. These are clearly more sophisticated psychological capacities than merely having conscious states.
(6) Zeki (2007) cites evidence that the "frontal cortex is engaged only when reportability is part of the conscious experience" (587), and "all human color imaging experiments have been unanimous in not showing any particular activation of the frontal lobes" (582).[3]
(7) Finally, another approach to human NCCs is by way of neurophysiological comparison to animals. The search for NCCs would seem to dictate that if a NCC is found in humans, then one would expect that the

corresponding area in animals would also underlie consciousness. This can also bolster support for the Animals Thesis, which I addressed in the previous chapter. Along these lines, Baars (2005) reviews significant evidence that conscious perception and cognition depend on the thalamocortical complex, which is also the basic neuroanatomy for many animals (including some nonmammals). Indeed, this is the main NCC area on Edelman and Tononi's view. Seth and Baars (2005) follow up on Edelman's earlier theory called "Neural Darwinism" (ND), which says that some groups of neurons are selected over others during brain development based on experience and behavior.[4]

I suppose it is possible that those who present the evidence just cited are just not looking at the PFC when recording their results. But I am not aware that these investigators are simply ignoring that part of the brain while trying to identify the NCCs in question. I do not mean to suggest that these findings are entirely uncontroversial. To be sure, there are those who argue that, based on other experimental results, PFC activity is implicated in having many first-order conscious states (Lau and Passingham 2006; Dehaene et al. 2006; Del Cul, Baillet, and Dehaene 2007; Gaillard et al. 2009). But once again, the main problem is that such experiments tend to demand explicit verbal reporting and introspection, which is not necessary for first-order conscious states and would involve neural structures that go well beyond the demands of HOT theory.

Regardless of the ultimate outcome of this disagreement, however, we can already see how well the HOT-Brain Thesis has been supported thus far. Indeed, adherence to any form of HOT theory provides an impetus behind a compelling and testable area of brain research. HOT theory can help to direct specific areas of brain research in a way that could cause either the modification or abandoning of HOT theory. In addition, a number of other scientists have found HOT theory useful in theorizing about consciousness (e.g. Weiskrantz 1997, Flohr 2000, Rolls 2004). Thus I strongly disagree with Revonsuo's claim that "these theories [= HOT theories] have not had any major impact on the empirical study of consciousness" (2010, 189). Not only have they already had some impact, but, as I hope to have shown more clearly by the end of this chapter, they have more to offer. Moreover, it is not clear that one can so neatly separate "philosophical" from "empirical" theories in the way that he does (Revonsuo 2010, chaps. 10 and 11).

In any case, one might then ask: Why think that unconscious HOTs can occur outside the PFC? What is the positive evidence for this? I will mainly rely on Newen and Vogeley 2003 and the references therein.

Let us assume first that HOTs can be understood as a form of self-consciousness, which seems reasonable. Unconscious HOTs might then be regarded as a kind of "pre-reflective" self-consciousness in ways we have seen in previous chapters. Newen and Vogeley (2003) distinguish five levels of self-consciousness ranging from "phenomenal self-acquaintance" and "conceptual self-consciousness" up to "iterative meta-representational self-consciousness." The majority of their paper is explicitly about the NCCs of what they call the "first-person perspective" (1PP) and the "egocentric reference frame." Citing numerous experiments, they point to various "neural signatures" of self-consciousness. The PFC is rarely mentioned and then usually only with regard to more sophisticated forms of self-consciousness. Other brain areas are much more prominently identified, such as the medial and inferior parietal cortices, the temporoparietal cortex, the posterior cingulate cortex, and the anterior cingulate cortex. Indeed, Damasio (1999) singles out the anterior cingulate cortex as a site for some higher-order mental activity or "maps."

There are also various cortical association areas that might be good candidates for HOTs depending on the modality. For example, key regions for spatial navigation comprise the medial parietal and right inferior parietal cortex, posterior cingulate cortex, and the hippocampus. Even when considering the neural signatures of theory of mind and mindreading, Newen and Vogeley cite and have replicated experiments indicating that such metarepresentation is best located in the anterior cingulate cortex. In addition, "the capacity for taking 1PP in such [theory of mind] contexts showed differential activation in the right temporo-parietal junction and the medial aspects of the superior parietal lobe" (538). Once again, even if the PFC is essential for having certain HOTs, this poses no threat to HOT theory provided that the HOTs in question are of the more sophisticated introspective variety.[5]

My conclusion thus far is that it is a mistake, both philosophically and neurophysiologically, to claim that HOT theory should treat first-order conscious states as *very* widely distributed in the brain, that is, as including PFC activity. If other HO theorists endorse a very global view, then so much the worse for them. However, to tie this together with the themes of the previous two chapters, I make the following concession:

If all HOTs occur in the PFC, *and if* PFC activity is necessary for all conscious experience, *and if* there is little or no PFC activity in infants and most animals, then either (a) infants and most animals do not have conscious experience or (b) HOT theory is false.

Unlike Carruthers (2000, 2005) and perhaps Rosenthal, I would opt for (b). I think I am more sure of animal and infant consciousness than any philosophical theory of consciousness. However, as we have seen in some detail, a good case can be made for the falsity of one or more of the conjuncts in the antecedent of the foregoing conditional.[6]

The NCC of HOTs need not include the PFC, and though HOT theory demands that conscious states be distributed to some degree, I opt for a more moderate global view, especially with respect to first-order conscious states. Thus, the total neural basis for first-order conscious states need not include the PFC.

So, contra Block, my view is that phenomenology does include at least some cognitive access, that is, a HOR of the conscious state. Any HOT theorist is committed to the view that a conscious state necessarily involves some higher-order cognitive access. Moreover, it seems best to treat the state as a combination of both the lower-order state and the HOR, as I have urged in arguing for the WIV in chapter 4. Thus the access consciousness in question is a part of the conscious state, not merely a causal contributor to it.

9.3 Parts, Wholes, and Feedback Loops

In this section, I examine in further neurological detail the notion that conscious states are complex states with HOTs (or METs) as part of the overall state. Recall that in chapter 4 I made several relatively brief observations regarding relevant neurophysiological evidence that I think supports the HOT theory in general and the WIV in particular.

9.3.1 More on Feedback Loops

We have seen that Edelman and Tononi (2000a, 2000b) argue that feedback loops (or reentrant pathways) in the neural circuitry of the brain are essential for conscious awareness. Churchland explains that "it is a general rule of cortical organization that forward-projecting neurons are matched by an equal or greater number of back-projecting neurons" (2002, 148–149). The brain structures involved in loops seem to resemble the structure of at least some form of HOT theory whereby LO and HO states are combining to produce conscious states. Recall that we also argued that the WIV is best supported by the evidence. In the WIV, essential and mutual interaction occurs between the relevant neuronal levels. Edelman and Tononi (2000a, 2000b) emphasize the global nature of conscious states, and it is reasonable to interpret this as the view that conscious states are composed of

both the higher- and lower-order components. They refer to the "dynamic core" as generally "spatially distributed and thus cannot be localized to a single place in the brain" (Edelman and Tononi 2000a, 146). However, they mainly locate reentrant cortical feedback loops in thalamocortical systems and not in the PFC (Edelman 1989; Edelman and Tononi 2000a, 2000b). So although conscious states are global in the sense that they cannot be localized to a small population of neural activity, they advocate a more moderate dynamic core. I agree. To further bolster this line of thought, consider the following:

(a) Bullier (2001) points to TMS studies that show that activation at the lowest cortical areas by feedback connection is necessary for conscious visual perception (cf. Pascual-Leone and Walsh 2001). For example, feedback from motion areas (MT/V5) to the primary visual area (V1) is necessary for visual awareness.
(b) Lamme (2003, 2004; see also Lamme and Roelfsema 2000) argues that recurrent processing is necessary before the properties of an object are attentively grouped and the stimulus can enter consciousness. Based on experimental results, such as texture segregation and visual search tasks, Lamme argues that the so-called feedforward sweep is not sufficient for consciousness. Like Malach and others, Lamme is careful to distinguish between consciousness and reportability. Although more cautious in tone, Pollen (1999, 2003) largely concurs with Lamme and presents similar data and rationale. If there is extensive damage to early visual areas (such as V1), then there will also be no conscious vision (such as in blindsight), but that is mainly because the process of conscious vision has been damaged at an even earlier stage.
(c) Lamme also explains that *backward masking* renders a visual stimulus invisible by presenting a second stimulus shortly after the first (about 40 msec later but perhaps up to 110 msec). Nonetheless the masked (invisible) stimulus still evokes significant feedforward activation in visual and even nonvisual areas. It seems that the feedback interaction from higher to lower visual areas is suppressed by backward masking, thereby disrupting reentrant processing (Fahrenfort, Scholte, and Lamme 2007; Kouider and Dehaene 2007).
(d) To use one nonvisual example, consider tactile awareness in the somatosensory cortex, extensively reviewed in Gallace and Spence 2010. Once again there seems to be evidence of feedback activity from higher brain areas, which is necessary for conscious tactile experiences. Gallace and Spence explain that "activation of early sensory areas is insufficient to sustain awareness of tactile sensations. . . . Higher order structures seem

necessary" (2010, 50). So tactile information becomes conscious when earlier somatosensory areas trigger a feedback signal from a higher-order representation.

Although there are some important differences among the more empirical theories described in this chapter, they do seem to converge on the notion that NCCs are at least somewhat widespread and involve feedback loops and integration between parts of the brain. Nonetheless PFC activity is usually not viewed as essential to first-order consciousness.[7] We can thus think of reentrant feedback as an unconscious HOT or MET directed at a lower-level mental state M. M will not become conscious without a MET. However, M may persist unconsciously during the feedforward sweep. During a first-order conscious perception, the MET itself is unconscious. Thus when researchers refer to feedback loops as "top-down attention" (Kouider and Dehaene 2007), we must keep in mind that such attention is not the conscious or voluntary kind.

So, as Kouider and Dehaene (2007) point out, there are two ways for an input stimulus not to reach consciousness, that is, two kinds of unconscious processes. There is first what they call *subliminal processing*, which is a "condition of information inaccessibility where the bottom-up . . . activation itself is insufficient to trigger large-scale reverberation" (869). For example, competition with other stimuli can prevent bottom-up strength of an initial stimulus. According to HOT theory, we might say that the subliminal unconscious occurs when M lacks the requisite strength to trigger a HOT. On the other hand, what they call *preconscious processing* "occurs when processing is limited by top-down access rather than bottom-up strength" (870). For example, backward masking and inattentional blindness temporarily prevents top-down access. In the HOT theory, we might say that preconscious processing occurs when, although M does have the requisite strength, a HOT is not formed due to a lack of top-down attention.

We must again be careful to be consistent with what has been said thus far with respect to subpersonal nonconceptual content (such as Marr's or Pylyshyn's theories) and to cases of cognitive impenetrability (such as experiencing the Müller-Lyer illusion). With regard to subpersonal nonconceptual content, there seem to be stages of, say, visual processing that are *very* early and best described as using concepts not possessed by a subject. This is the sort of unconscious processing closer to what Kouider and Dehaene (2007) have in mind by subliminal processing. Such information is not even potentially conscious. Other unconscious processes, however, are indeed potentially conscious and become so when accompanied by the appropriate feedback loop. With regard to cases of cognitive impenetrability,

we may also suppose that there are occasions where higher-level conceptual content is unable to influence or alter some aspects of a resulting visual experience because of how early in visual processing it occurs. Indeed, it seems that some very early visual processing, say, 70 msec or less, is immune to cognitive influences.

In chapter 4, I also alluded to work of Feinberg, who offers a helpful way to think about this issue (Feinberg 2000, 2001, 2009). He argues for the nested hierarchy theory of consciousness (NHTC), which has some affinities to HOT theory and the WIV. In a nonnested hierarchy, lower and higher levels are independent entities in which the top of the hierarchy does not overlap with the bottom. A nonnested hierarchy has a pyramidal structure with a clear top and bottom with the higher levels controlling the lower levels analogous to a military command structure. In a nested (or compositional) hierarchy, lower levels of the hierarchy are nested within higher levels to create increasingly complex wholes. Unlike some accounts of neural hierarchy, which view the brain as a nonnested hierarchy, the NHTC (like the WIV) would treat some areas of the brain as a nested hierarchy when conscious states occur. The idea is that lower-order features combine in consciousness as *part of* (or nested within) higher-order features. So consciousness is not narrowly localizable, but it is also not exceedingly global. Conscious states are thus neurally realized as combinations of lower- and higher-order brain features. So, as I argued in earlier chapters, we might view the WIV as a case where a conscious mental state is a complex of two parts that are integrated in a certain way. Like the NHTC, there is essential reciprocity between cortical and subcortical structures. The structures in question are not merely laid on one another without neural functioning going in both directions.

Recall also that the WIV is best thought of as an *interactive* theory such that "once a stimulus is presented, feedforward signals travel up the visual hierarchy. . . . But this feedforward activity is not enough for consciousness. . . . High-level areas must send feedback signals back to lower-level areas . . . so that neural activity returns in full circle" (Baars and Gage 2010, 173). Higher areas need to check the signals in early areas and confirm if they are getting the right message. Such recognition or "re-cognition," as we called it, is essential for conscious states.

Kosslyn offers a similar hypothesis with respect to visual consciousness. According to him, consciousness "is not simply a byproduct of activation of any one area, or even a set or areas, but rather occurs when processing going downstream must mesh with processing going upstream" (2001, 91). Thus, given the plethora of feedback connections, it is reasonable to suppose that

"consciousness arises at junctures where different types of representations meet and feedback and feedforward flows must be coordinated" (97). For example, the outputs from retinotopically visual areas, where the pattern of stimulation on the retina is projected onto the cortical surface of these areas, preserving much of the spatial layout of the stimulus, mesh with top-down feedback from areas via back projections. We might say, therefore, that we become aware of the lower-level retinotopic areas during this process. A similar view would likely hold for other sensory modalities, such as the auditory and somatosensory areas. Each area has similar or analogous cytoarchitecture, that is, the distribution of neurons in each cortical layer (Grossenbacher 2001).

With respect to the WIV, then, I appealed to the mereological notion of an "underlap" in the sense that when M (x) underlaps MET (y) there is a CMS (z) such that M is part of CMS, and MET is part of CMS. However, there can also still be some *overlap* between M and MET insofar as there is a psychologically real relation between M and MET. On the neural level, there are also what seem to be overlapping areas of feedforward and feedback loops between M to MET.

Thus if we are genuinely interested in the NCC for a *first-order* conscious state, we should be prepared, as Koch explains, to find the minimal set of neuronal events and mechanisms jointly sufficient for a specific conscious percept. In my view, the NCCs for such states do not include the PFC but, depending on the sensory modality, would include various other higher-association brain areas.

I have also suggested that the importance of higher-order *concepts* in conscious experience is apparent and essential. This is related to my quasi-Kantian interpretation of the interaction between the sensibility and the understanding. Moreover, it fits well with the thesis of conceptualism, which I defended at length in chapters 6 and 7. It is clear that part of the reason why Edelman and others believe that back projections play a prominent role in consciousness is that, as Churchland puts it, "perception *always* involves classification; conscious seeing is *seeing as*" (2002, 149). If the WIV is true, then we should expect to find that the neural realization of METs interacts with the neural realization of lower-order mental states. Moreover, if concepts constitute thoughts, then it stands to reason that the constituents of thoughts would be localized within the neural realization of thoughts. The neurological basis of concepts is, in some ways, even less clear than NCCs. Some have offered interesting hypotheses regarding the *other* NCCs, that is, what we might call the "neural correlates of *concepts*" (Miller et al. 2003), as well as the neurobiology of category learning (Ashby

and Spiering 2004; Ashby and Maddox 2005; Kéri 2003). Weiskopf (2007) refers to the "NCCT," that is, the neural correlates of *conceptual thought*.

For my purposes, one potentially promising area of research has to do with "convergence zones" (Damasio et al. 2004). In the context of critiquing Prinz's (2002) neoempiricist account, Weiskopf explains that convergence zones are "neural ensembles that receive projections from earlier cortical regions (e.g., lower-sensory areas), contain feedback projections on those earlier layers, and feed activity forward into the next highest layers of processing. These zones can also refer back to zones earlier in the processing stream, so that higher-order zones may be reciprocally connected with multiple lower-order zones. The functional role of convergence zones is to orchestrate the reactivation of lower-order activity patterns" (2007, 162). This description seems tailor made for the WIV and the role of METs. Convergence zones are not PFC areas, though some do receive input back from the PFC. Convergence zones integrate two or more brain areas with what seem to be the kind of reciprocal interaction demanded by the WIV. They also receive downward connections from the cingulate cortex, basal ganglia, and thalamus. Weiskopf also explains how there must be "content matching" for the "appropriate perceptual representations to be retrieved from memory and compared" (180). In almost WIV-like terminology, he tells us that "occurrent perceptual representations are causally intertwined with activity in numerous non-perceptual regions" (181). Again, it is not a problem for my theory that *some* METs occur in the PFC as long as they are more sophisticated introspective states having to do with conscious inference, reasoning, executive control, language use, reflective consciousness, and so on.

Convergence zones "refer back" to lower areas, suggesting a representational relation, Weiskopf argues, not merely nonrepresentational mechanisms. This is consistent with METs as one part of a conscious state. Weiskopf makes a compelling case that it is simply not easy, neurologically speaking, to separate out perceptual contents from conceptual contents. This is precisely what the WIV predicts, not to mention the Kantian-style defense of conceptualism defended earlier. Moreover, this account is consistent with the theory of content presented earlier in this work. Nonperceptual representations are reliably caused by lower brain areas and play a central role in the categorization process. So, according to Rupert's theory, repeated exposure to certain stimuli will cause not only the lower-level representations but also higher-level representations. Recall that content and concept acquisition is determined by past relative frequency (PRF) and a substantive developmental process. A mental representation R "has as

its extension the members of natural kind K if and only if members of K are more efficient in their causing of [R] in S than are the members of any other natural kind" (Rupert 1999, 323). The notion of "efficiency" is cast in terms of numerical comparisons between the PRFs of certain causal interactions. But the same reasoning should also apply to representations of mental states, such as a MET representing M. Indeed, on the neural level, some scientists speak of changes in neural connections, as well as strengthening those connections over time.[8]

We must also keep in mind that much of the literature on concept learning is concerned with sophisticated instances of explicit or conscious learning, as opposed to implicit learning, depending on the experimental task. As we have seen, the notion of concept *acquisition* is more general and less sophisticated, for example, in artificial grammar learning, the dot pattern task, and implicit learning in amnesiacs (chap. 7). There seems to be some consensus that several different brain areas are involved in detecting similarity, discrimination, and recognition, including the basal ganglia, sensory neocortex, and the medial temporal lobe. They may also be subsystems that feed into the PFC depending on the difficulty of the task at hand. But this is as it should be. When there is an introspective state, a conscious MET, there should then be a yet-higher-order (third-order) MET directed at a MET.[9]

9.3.2 A Connectionist Approach?

Finally, it might occur to some readers that I am endorsing what could be construed as a connectionist approach to conscious states in the sense that there are patterns of neural activity augmented by backpropagation and resulting in concept learning and application (Rumelhart and McClelland 1986). This is indeed an intriguing idea in connection with the WIV, but we must be careful not to overstate the similarities. First, a brief description of connectionist networks is in order (Garson 2010; Waskan 2010).

Connectionism is an artificial-intelligence approach to the study of human cognition that hopes to explain human intellectual abilities using artificial *neural networks* (or "neural nets"). This approach is also sometimes called Parallel Distributed Processing (PDP). Neural networks are simplified models of the brain composed of large numbers of neuronlike units, together with *weights* that measure the strength of connections between the units. These weights model the effects of the synapses that link one neuron to another, as I described earlier in the chapter. Units in a net are normally grouped into three classes: *input units*, which receive information to be processed, *output units*, where the results of the processing are found, and units in between called *hidden units*. Experiments on these models have

demonstrated an ability to learn skills such as face recognition, reading, and simple grammatical structure. One influential early connectionist model was a net trained by Rumelhart and McClelland (1986) to predict the past tense of English verbs. The pattern of activation set up by a net is determined by the weights.

Now, when activation flows directly from inputs to hidden units and then on to the output units, this is called a "feedforward net." It is well understood that a truly realistic model of the brain would have to include many more layers of hidden units, as well as recurrent connections that send signals back from higher to lower levels. However, one of the most widely used training methods in PDP is called "backpropagation." To use this method, one needs a training set consisting of many examples of inputs and their desired outputs for a given task. This allows for backpropagation learning, where an error signal propagates backward through multiple layers to guide weight modifications. Finding the right set of weights to accomplish a given task is the main goal in connectionist research.

There are indeed some similarities between the WIV and neural nets. First, perhaps most obvious is the notion of backpropagation, which has a clear analogue of feedback loops in the WIV. Second, there is the notion that, through increasingly strengthened connections, a system can ideally learn concepts and acquire mental content in a way similar to the Rupert-style view first introduced in chapter 2. Third, with regard to concept acquisition, we saw how important implicit (unconscious) learning was to making sense of the WIV and conceptualism while avoiding radical nativism.

In line with the foregoing points, perhaps most interesting to me is recent work by Cleeremans, Timmermans, and Pasquali (2007), who explicitly link connectionism to consciousness and metarepresentation. In essence, they present results of two simulations designed to show how a limited form of metarepresentation or "re-representation" can be realized in a connectionist network. A first-order network is trained to perform a categorization task that, in turn, becomes input for a second-order network, so that the second-order network "observes" the first-order network. The second-order network can be trained either simply to copy or "encode" that previous output or to perform a further task, such as evaluating the first-order network's performance. The overall suggestion, much like the WIV, is that consciousness involves recognition or awareness of lower-order states, that is, a mind understanding its own workings. Consciousness results not merely when one learns about the world but also when the mind learns about itself. Conscious experience occurs when an information-processing system has learned about its own representations of the world.

With regard to learning, then, Cleeremans, Timmermans, and Pasquali (2007, 1035) explain that "at some point during the early stages of learning, some aspects of the learned knowledge become available as targets of higher-order representations. In other words, whereas initially unstable first-order knowledge makes it impossible for the higher-order network to consistently learn about them, this changes with training in such a manner that once first-order representations have become sufficiently stable, the higher-order network can then use the structure that they contain so as to improve its own ability to reconstruct the input and the output of the first-order network successfully." They also point out how "learning might initially take place in an essentially implicit [or unconscious] manner" (ibid., 1037) and so perhaps similar to TILT. The claim is that the metarepresentations are learned in the same automatic or unconscious way as first-order representations. Conscious mental states are thus the result of unconsciously learned lower-order representations accompanied by unconsciously learned metarepresentations. This is much like the overall theory presented in this work.

Several others have also explicitly tied together connectionism with implicit learning, early development, and consciousness (Cleeremans and Jiménez 2002; Mareschal 2003). For example, Mareschal (2003) looks to connectionist architecture for a possible explanation of infants' preferential looking techniques. Infants look longer when there is a discrepancy between stored information and input from a stimulus. While attending to the stimulus, the infant updates and adjusts its internal representations. When there is no longer such a discrepancy, such as when viewing a familiar object, attention is switched elsewhere. Without embracing a full-blown theoretical understanding of concepts, it also makes sense to suppose that any innate or very early infant concepts can serve as an initial state of an organism's mind.[10]

Despite these exciting lines of thought, we can also identify some clear differences between the WIV or HOT theory and connectionist accounts (not to mention the usual dissimilarities between a real brain and a connectionist network). Most important perhaps is the well-known dispute about whether or not connectionist networks are genuinely "representational," at least with regard to compositionality (productivity) and systematicity (Fodor and Pylyshyn 1988). In short, the problem is that connectionist architecture cannot account for essential aspects of thought, such as the ability to think many thoughts by simply recombining or reordering their concepts, which, in turn, requires the systematicity of syntax and semantics. Like language, the productivity and systematicity of thought, as well as reasoning and inference, are explained by its combinatorial and recursive

syntax and semantics. If one can think "John loves Mary," then one can think "Mary loves John." There is at least no guarantee of systematicity for a given network. Moreover, if concepts and thus thoughts are distributed states of networks, it begins to look as if they are not explicitly represented at all in the connectionist units.[11]

Nonetheless some have forcefully argued that connectionist networks are compatible with classical models of mental representation and can make sense of the requisite compositionality (Hawthorne 1989; Chalmers 1993). It is not clear to me that connectionist models are inconsistent with classical models, but I will not enter into this debate here. Obviously, however, if any theory of human cognition does not allow for genuine mental representation, then it cannot realize a *higher-order* representational theory, and I would be inclined to reject it. As we have seen earlier in the chapter, however, we seem to be able to make sense of degrees of "distributed" representation. Few, if any, today hold that there are extremely localized mental representations in the brain.

Overall, it is starting to seem as if many of the pieces of the puzzle are coming together fairly nicely. However, I now turn to yet another difficult problem.

9.4 The Binding Problem and the Unity of Consciousness

In this final section, I briefly explore the well-known binding problem and the interrelated topic of the unity of consciousness (Dainton 2000; Cleeremans 2003).[12] Indeed, as we shall see, these problems are importantly related to the search for NCCs and perhaps even the hard problem.

9.4.1 The Problem

One important aspect of conscious experience is that it seems to be "unified" in at least one important sense. This crucial feature of consciousness played an important role in the philosophy of Kant, who, as we have seen, argued that unified conscious experience must be the product of the (presupposed) synthesizing work of the mind, including the application of various concepts or "categories." Kant famously called the activity of such synthesizing (or binding) the "transcendental unity of apperception." Although Kant had nothing to say about such mechanisms in specifically neurophysiological terms, he had much of value to say about them in cognitive terms (Brook 1994, 2005; Kitcher 1990, 2010). I have also drawn on Kant in developing the WIV, especially in regard to the synthesis and integration of the perceptual and conceptual.

Getting clear about exactly what is meant by the "unity of consciousness" and explaining how the brain achieves such unity has become a central topic in consciousness studies. There are many different senses of "unity" (Tye 2003; Bayne and Chalmers 2003; Brook and Raymont 2010), but perhaps most common is the notion that, from the first-person point of view, we experience the world in an integrated way and as a single phenomenal field of experience (Cleeremans 2003). However, when one looks on the neural level at how the brain processes information, one sees only complex discrete regions of the cortex processing separate aspects of perceptual objects. Even different aspects of the same object, such as its color, shape, and motion, are processed in different parts of the brain. When a blue ball is thrown, we visually experience the motion, color, shape, and object all at the same time; it is not as if these properties come apart in our visual experience. So properties also become unified in objects. One can consciously experience the ball and even catch or hit the ball without consciously attending to the ball's properties.

Given that there is no "Cartesian theater" in the brain where all this information comes together (Dennett 1991), the problem arises as to how the resulting conscious experience is unified. What mechanisms allow us to experience the world in such a unified way? How can such unity arise from such diversity? What binds together such disparate neural activity to produce the kind of unity we experience from the first-person point of view? "The problem of integrating the information processed by different regions of the brain is known as the binding problem" (Cleeremans 2003, 1). And what happens when this unity breaks down in various pathological cases?

Bayne and Chalmers (2003) attempt to clarify what is meant by the unity of consciousness, and a number of important interconnected theses emerge. Perhaps most central is what they call the "unity thesis," according to which "necessarily, any set of conscious states of a subject at a time is unified" (24). Shoemaker (2003) argues that consciousness requires the unity of consciousness. He stresses the need to accept "consciousness holism" (as opposed to "consciousness atomism"), whereby a mental state's phenomenal character depends, in part, on other "co-conscious" states. I return to their views in section 9.4.3.

9.4.2 A Few Theories of Binding

In addition to Crick and Koch's temporal synchrony account, perhaps best known is Anne Treisman's (1993, 2003) theory that distinguishes three forms of binding: properties, parts, and perceptual grouping. Her influential "feature integration theory" (FIT) emphasizes the role that spatial attention

plays in selecting the appropriate features to be bound. Relating her theory to the temporal synchrony account, she points out that the two are consistent, but also that "the binding problem is really two separate problems: how do we select the correct combinations of features to bind, and how are their conjunctions encoded and maintained once they have been bound? The spatial attention account offered by FIT answers the first, whereas the synchronized firing account deals with the second" (2003, 104–105). Binding is essential for conscious experience, but it is important to be clear about the way she uses the notion of "attention" (see chap. 6). Obviously much of what Treisman has in mind is unconscious. As we have seen, this is perfectly consistent with the WIV with its emphasis on unconsciously applied concepts presupposed in the resulting experience. Some binding seems to occur early in visual processing, but some also occurs as a result of applying concepts via a MET. Either way, however, conscious attention is not required, and Treisman's account seems consistent with the WIV.[13]

It is also worth mentioning two other neurophysiological accounts found in Cleeremans 2003, offered primarily as alternatives to the temporal synchrony view.

(1) O'Reilly, Busby, and Soto (2003) argue that the binding problem should be addressed differently in different areas of the brain: the generic cortex, hippocampus, and prefrontal cortex. In very much a connectionist spirit, they discuss some limitations of the temporal synchrony view, such as the difficulty in accounting for its "fragility." It would seem that having the unity of consciousness depend so heavily on precise timing relationships implies that any interference would drastically affect our consciousness. But this fragility contradicts the "evidence that in fact our brains are highly robust and subject to rather graceful degradation" (172). One advantage of connectionist architectures is precisely this notion that damage or deterioration is gradual or piecemeal, rather than a complete shutdown or system crash. Moreover, the notion that there are multiple locations where binding takes place is certainly consistent with the WIV. We might say that there are mini-bindings that occur at various levels in the brain, from early visual or auditory areas to various higher cortical areas.

(2) Humphreys (2003) echoes this view and offers evidence that binding is not a unitary process but instead involves multiples stages that can become dissociated from each other as a result of brain injury. For example, he explains that people with achromatopsia and integrative agnosia give us reason to believe that "visual processing is fractionated at a neural level, with different regions specialized for coding color, motion, form, and location information" (115). A subject with achromatopsia can have selective

loss of color vision without having an impairment in motion vision. Likewise, patients with integrative agnosia, among other things, show that we bind locally oriented elements into edges before binding edge elements into holistic shapes. As we saw in chapter 6, associative agnosia could be viewed as involving a breakdown in the unity of consciousness. Subjects are unable to understand what they are seeing in the sense of getting the big picture, applying the proper concept, and putting the parts of an object together so as to experience the whole. Thus Humphreys rejects what he calls "one-shot" accounts of binding, such as temporal synchrony, and argues that the "unity of consciousness derives from several separable neural processes of binding" (114). However, it is important to note that there could still be a single *mechanism* of binding even if binding occurs at separable stages finally resulting in unified conscious perception.[14]

It has long been interesting to examine what happens in abnormal cases where unity breaks down. Young (2003), for example, discusses prosopagnosia, the inability to recognize familiar faces overtly, and the Capgras delusion, the belief that other people, usually close relatives, have been replaced by imposters. In addition to an abundance of empirical data showing that prosopagnosics exhibit some covert recognition of familiar faces, Young observes that "the Capgras delusion might form a kind of mirror image of prosopagnosia. In prosopagnosia, overt recognition is impaired but emotional orientating responses may be relatively preserved. In Capgras delusion, overt recognition is relatively preserved but emotional orientating responses are lost" (243). There are many different pathologies of interest to philosophers and psychologists alike, such as "split brain" or commissurotomy cases, amnesia, dissociative identity disorder (formerly called multiple personality disorder), and schizophrenia. Patients with Bálint's syndrome (or simultanagnosia) see only one object at a time located at one "place" in the visual field. Subjects seem not to be aware of even two items or objects in a single overall conscious state. Each of these disorders is or could be the subject of a separate chapter or book.[15]

9.4.3 Unity and the HOT Theory

These neurophysiological accounts of binding seem consistent with HOT theory and the WIV, and there is no point in becoming unnecessarily wedded to any one of them. Indeed, they may not all be mutually inconsistent. The main type of unity discussed in the previous paragraphs has to do with what Bayne and Chalmers call "objectual unity": "Two states of consciousness are *objectually unified* when they are directed at the same object" (2003, 24). The earlier case of seeing a single moving (round) blue ball would be

an example. It would seem that such binding (or synthesizing) takes place unconsciously and is simply presupposed in the resulting experience of an object. A HOT theorist can explain that whatever the ultimate neurophysiological story, the resulting conscious experience's content includes experienced properties of the object and is also included in a HOT's contents. Once again, a HOT is not conscious when one has an outer-directed conscious visual perception.

A somewhat different notion of unity arises when we consider whether or not a *conjunction* of conscious states yields a *further* conscious state. This is closer to what Bayne and Chalmers call "subsumptive unity": "Two conscious states are *subsumptively unified* when they are both subsumed by a single state of consciousness" (2003, 27). For example, we might suppose that auditory and visual conscious states can combine into an overall conscious perceptual state. In such cases, we might say that several different modality-specific HOTs occur at approximately the same time. The result is that the subject experiences the conjunction of the states, such as hearing and seeing a band at a concert, or feeling and hearing a wave crashing onto the beach, and so on.

It is not clear to me, however, that for there to be a unified experience, there must be some single, all-encompassing conjunctive HOT *in addition to* the individual modality-specific HOTs. Each HOT would normally reference the same "I" to which those states belong. Bayne and Chalmers do suggest that, in addition to each conscious state, there is *also* the overall conscious perceptual state. "What is important . . . is that this total state is not *just* a conjunction of conscious states. It is also a conscious state in its own right" (27). This ultimately gives rise to what they call the "total conjunctive unity thesis," which states that "if C is the conjunction of all of a subject's phenomenal states at a time, then C is itself a phenomenal state" (46). Conversely, it is also unclear that even if one does have a single, all-encompassing conjunctive conscious state, there are also a number of simpler conscious states. In addition, we have seen that further complications arise once we allow that consciousness can admit of degrees, such as in peripheral and focal consciousness. Indeed, Bayne and Chalmers are much more cautious later in their essay, when they say that thinking in terms of "a mereological part/whole relation among phenomenal states" should be viewed more as an "aid to intuition rather than a serious ontological proposal" (40).

Recall also the related "complexity problem" discussed in connection with conceptualism (in chap. 6). For example, let us suppose that one has a conscious state that combines the visual image of the waves on a beach,

the feel of the water, and the sound of the waves crashing. In these multisensory cases, I urged that it is wiser to hold that there are several HOTs, at least one for each sensory modality, which are bound together to produce a complex experience. But how does this happen? On the cognitive side, we might simply say that there is a "subject unity" that binds these HOTs and thus conscious states together. The reference of "I" in each HOT implicitly refers to the same subject of conscious experience. I will return to this issue shortly.

As Brook and Raymont (2010, sec. 6) nicely explain, one way to frame some of the disagreement is in terms of those who favor the "experiential parts" view (EP) as opposed to the "no experiential parts" view (NEP). The EP view says that unified conscious experience includes simpler experiences as parts or something like parts. The NEP view is basically that the conscious state through which diverse contents are presented does not have other conscious states, experiences, as parts. According to the EP theory, I am conscious of many experiences when I have a unified conscious experience (Dainton 2000; Bayne and Chalmers 2003; Shoemaker 2003). According to the NEP view, I am conscious of just "one experience" with many different contents (James 1890; Searle 2002; Tye 2003). The unified composite experience *replaces* or *supersedes* any included conscious states. I am inclined to favor the NEP theory for two main reasons, though I do not think that HOT theory is logically committed to it.

First, with respect to experienced objectual unity, it does not normally seem to be the case that we can separate out the shape, color, and movement experiences of the ball. Phenomenologically, we experience the object and its properties all at once. Second, and perhaps more important, the EP theory seems to undermine the main point behind the very binding problem itself. After all, isn't the binding problem generated precisely *because*, from the first-person point of view, we only experience all the object's properties *together* in a single visual experience? It is *logically* true that if I consciously experience an object O that has properties A, B, and C, then I am experiencing A, B, and C. But this logical entailment does not mean that I have three *further* distinct conscious *experiences* over and above my experience of A, B, and C all at once. In addition, some individual HOTs may not be complex at all because they refer only to objects or properties in one's peripheral awareness. As we saw in chapter 6, this also partly explains away the alleged richness of conscious experience.[16]

One might object that I am now conceding that there are composite HOTs after having rejected them earlier. But this is not the case; we are now discussing a composite *experience* with complex contents that partly results

from many different unconscious HOTs. In normal cases, one's conscious experience is indeed unified. But when there is a composite experience, there are not further experiential parts.

Notice that the kind of composite experiential state at issue is different from the sort of composite I described in terms of the WIV (in chap. 4), where two *unconscious* but interrelated states combine to form a conscious state, provided that certain conditions are met. This is entirely different from claiming that there are two conscious states combining to form a separable third combined conscious state. We have also seen that simply because a conscious *experience* has multiple *contents* (or even attitudes), it does not follow that there are also multiple conscious experiences or vehicles. Indeed, I argued in chapter 4 that a single conscious state can have multiple contents and attitudes. Individuating conscious states in terms of attitudes, let alone contents, is not the only plausible alternative. Moreover, as we saw with respect to the unconscious parts objection in chapter 4, one might even wonder if the EP view and the total conjunctive unity thesis represent an instance of the fallacy of composition. On the flip side, perhaps the total subsumptive unity thesis is a case of the fallacy of division.

Now, in terms of the underlying neural realization of subsumptive and conjunctive unity, perhaps there are convergence zones where more than one conscious state is "tied together" in some way. But it would seem unreasonable, as Bayne and Chalmers recognize, to insist that there is a place in the brain where a conjunctive state *all* comes together in addition to the neural activity in different brain areas. This would especially seem unlikely when a conscious state is multimodal. Indeed, even *within*, say, the visual cortex there is little reason to suppose that there is a *separate* conjunctive state for motion, color, and shape. Again, this is what mainly generates the binding problem in the first place. Moreover, even if an overall brain state comprises many neural parts, it does not follow that each part is itself a conscious state.

Some authors have argued that HOT theory is fundamentally at odds with the unity thesis. For example, Bayne and Chalmers (2003, 50–53) argue that, according to HOT theory, we have little reason to think that several HOTs at a given time must result in a conjunctive HOT. But in light of the earlier discussion, we can now see how to reply. We can concede the point to Bayne and Chalmers, but then ask: why is it a problem for HOT theory specifically? The unity of consciousness results more from a sense of *subject unity*, not from an explicitly distinct conjunctive content of HOTs. In cases of disunity or psychopathology, such as achromatopsia or

akinetopsia, there is good reason to suppose that one of the typical HOTs is absent (or mismatched). Bayne and Chalmers (2003, 52–53) do recognize that HOT theory has several options at this point, such as denying or limiting the unity thesis. Limiting the unity thesis to typical or normal cases, for example, seems like a reasonable option. After all, we do find abnormal cases where unity breaks down. Not only should HOT theory allow for them, but it can also explain them by pointing out how HOTs differ in those unusual cases (recall also the discussion of agnosia in chapter 6). Thus, if confronted with the choice between the unity thesis and HOT theory, I would opt for HOT theory.

Shoemaker (2003, 60–64) argues that HOT theory is too "atomistic," which is the view that what makes a state conscious is independent of the factors that make two or more states unified. A particular state's consciousness depends on properties associated only with that state, as HOT theory says. So, according to Shoemaker, HOT theory is thus inconsistent with the unity thesis, and HOT theory is false. He holds a holistic view that says that consciousness requires "co-consciousness" and thus requires the unity of consciousness.

But, first, we have already seen how HOT theory can explain the unity of consciousness, or at least the appearance of unity, in terms of subject unity (see also below). Second, Shoemaker offers a weak argument against HOT theory based on an alleged counterexample whereby a HOT is directed at a clearly unconscious mental state, a repressed shameful wish. Shoemaker then imagines that the HOT is also repressed. But, as I explained in chapter 2, HOT theory can avoid these kinds of arguments by invoking the noninferentiality condition, namely, that the HOT in question cannot arise via inference. If it does, then no conscious state will occur. Shoemaker does not provide enough detail in his thought experiment to be sure, but it is difficult to see how else one would come to possess a HOT directed at one's own repressed shameful wish. It is also unclear why Shoemaker thinks that *both* the unconscious mental state and the HOT are not even access conscious. Repressed mental states are not necessarily forever inaccessible. Given the right circumstances and perhaps some therapy, they can be brought to consciousness. On the other hand, if my lengthy defense of the HOT Thesis is plausible, then a HOT directed at a repressed shameful wish would become conscious.

Finally, Dainton refers to what he calls "the HOT Unity Thesis," which says that "a collection of experiences at a given time t are phenomenally unified if and only if they are *all* the objects of *a* higher-order thought" (2007, 213; italics mine).

But first Dainton replies that we don't need to be (consciously) thinking about an experience for that experience to be conscious. Fortunately, he immediately correctly recognizes that this is not a problem for HOT theory, since the HOTs in question need not themselves be conscious. But then he simply states that "it is difficult to see how . . . HOT theory can hope to shed light on the nature of phenomenal unity viewed as a real and occurrent feature of consciousness" (213). Needless to say, I think that the extended argument of the present work and chapter go a long way toward shedding at least some light on this matter. There are occurrent HOTs with explicit reference to oneself. Finally, we need not adopt Dainton's HOT Unity Thesis as it stands. Instead we should hold that "when one has a unified conscious experience at time t there will be one or more HOTs directed at one or more LO states." Or "A collection of experiences at a given time t is phenomenally unified if and only if those experiences are *each* objects of higher-order thoughts." There is no need to assume that something like EP or the conjunctive unity thesis is true, nor must we suppose that a single, all-encompassing HOT exists when one has a unified complex experience.

There is also the more traditional side of this debate, which centers on the nature of the self and the problem of personal identity. This mainly arises when Bayne and Chalmers mention "subject unity"; namely, "Two conscious states are *subject unified* when they are had by the same subject at the same time. So all of my current experiences . . . are subject unified, simply because they are all *my* experiences" (2003, 26). As Bayne and Chalmers point out, the subject unity thesis runs the risk of triviality. However, HOT theory has something substantial to offer by way of explaining subject unity, namely, that an I-concept can be found in each HOT.

We might also think of subject unity as related to the traditional problem of whether or not there is a unified self in the way that Descartes assumed there to be. But HOT theory need not settle this deeper metaphysical problem to explain the unity of experience. I largely agree with Rosenthal (2003) that HOT theory can help to explain a traditionally important *sense of unity*, that is, why it at least *seems* to us that we are each a unified self. Unlike other theories of consciousness, self-reference through the self-concept "I" accompanies each and every conscious state. Rosenthal explains that "each HOT refers not only to such a [mental] state, but also to oneself as the individual that's in that state. This reference to oneself is unavoidable" (2003, 330). Explaining state consciousness requires reference to a subject of those conscious states. As we have seen, however, there are degrees of I-concept, some of which need not be sophisticated. And even when one is having a first-order conscious state, there is at least an unconscious HOT

with a self-referential component. Thus we again see the natural advantage of HOT theory over other accounts. HOT theory can explain why there is, or at least normally appears to be, unity of consciousness in a way that other theories of consciousness cannot.

Moreover, when we are engaged in introspection, the sense of "I" is increased because one is now *consciously* thinking that "I am in a mental state." "So introspecting our mental states results in a conscious sense of unity among those states even when the states are conscious by way of distinct HOTs" (Rosenthal 2003, 332). All of this leads us naturally to think, or perhaps merely assume, that there is a single unified self behind the appearances. Moreover, there are often relations between one's conscious states that add to one's sense of unity, such as overlapping content, memories, and making conscious inferences. One way to explain these phenomena is to posit that conscious states involving such relations at least normally belong to the same self or person.[17] Bayne (2004) also appeals to "co-ownership" and subject unity to explain the unity of consciousness. Although he holds the EP view, it also seems possible to favor the NEP view while relying on subject unity and HOT theory.

One may remain skeptical about the very existence of a "self" in the tradition of Hume, Kant, and, more recently, Dennett. But this is not a problem unique to HOT theory. We could certainly be wrong in thinking that there is a self to which all individual "I's" refer, especially if one means some stable substance in addition to conscious states. Perhaps there is no self, and it is a mere illusion. But, as we saw earlier, it is arguably most important to explain the *appearance* of the same "I." What matters most is the subjective impression of unity, not actual unity. The "aim here is not to sustain the idea that a single, unified self actually exists, but to explain our compelling intuition that it does" (Rosenthal 2003, 337–338). So the foregoing line of argument is not necessarily inconsistent with, say, Dennett's view that the self is merely a construct based on a linguistic "narrative." Similarly, Hume's so-called bundle theory, whereby a "person" is at best a bundle of sensations or perceptions, cannot perhaps be ruled out. But Hume's problem partly arises from "his tacit adoption of a specifically perceptual model of introspection; one cannot find a self when one seeks it perceptually. The HOT model, by contrast, provides an informative explanation of the way we do seem to be introspectively conscious of the self" (Rosenthal 2003, 333). At the least, perhaps Kant was correct that we cannot *know* that there is a self or what it is like. Nonetheless he recognized that the "I think" must be able to accompany all of one's representations or mental states (1871/1965, B131–B132). He also argued that we must distinguish between

any experienced or empirical self and the transcendental "I." Thus there is still a sense in which some form of self-consciousness is presupposed in consciousness.[18]

In any case, we have seen in this chapter how the idea that feedback loops in the neural circuitry of the brain are necessary for conscious awareness can be usefully adopted by HOT theory and especially the WIV. There are back-projecting neurons often matching the number of forward-projecting neurons, and these loops are essential for consciousness. In some ways, Edelman and Tononi are trying to explain the unity of consciousness in neurophysiological terms. Perhaps this account can complement the temporal synchrony approach. One line of future research would be to examine the extent to which the *combination* of temporal synchrony and feedback loops could be *sufficient* for consciousness and thus for the NCC. Like temporal synchrony, loops are ubiquitous in the brain, and such neural feedback activity can occur unconsciously. But we have seen how the properly located loops can also be sufficient for consciousness.

Finally, all of the discussion in this chapter could be related more directly to the hard problem understood as a neurophysiological problem, namely, just *how* or *why* subjective experience arises from the brain. I have already addressed this issue in chapter 4 from the perspective of a HOT theorist, but in my view, advances in one of the areas discussed in this chapter will also have a major impact on the others. Perhaps somewhat speculatively, we might suppose that there is really only one problem under three different names: the hard problem, the binding problem, and the neural correlates of consciousness. Discovery of the true mechanism(s) of binding, for example, might offer insights into the true NCCs and thus lead to solving the hard problem defined in this way. It is also clear that the binding problem is inextricably linked to explaining the unity of consciousness. As we saw, some attempts to solve the binding problem have just as much to do with isolating the precise brain mechanisms responsible for consciousness. For example, Crick and Koch's idea that synchronous neural firings are (at least) necessary for consciousness can also be viewed as an attempt to explain how disparate neural networks bind together separate pieces of information to produce unified subjective conscious experience. In addition, perhaps the explanatory gap between third-person scientific knowledge and first-person unified conscious experience can also be bridged. This exciting area of inquiry is central to some of the deepest questions in the philosophical and scientific exploration of consciousness.

To close this chapter, I think I have shown that the HOT-Brain Thesis is true. It says that there is a plausible account of how my version of HOT

theory might be realized in the brain and can lead to an informative neurophysiological research agenda. Alternatively, HOT theory is related to, and consistent with, a number of leading empirical theories of consciousness. It would be imprudent to rule out the idea that future evidence from neurophysiology might force me to rethink the relationship between HOT theory and the brain. I may eventually either abandon or modify the WIV in some way, depending on future research. However, I do not think that the WIV must be wedded to Edelman and Tononi's theory, but I think that it currently fits best with the WIV.

9.5 Conclusion of the Book

Overall, then, I believe that I have made a plausible case for each of the theses presented in chapter 1 that comprise what I termed the Consciousness Paradox. Perhaps equally important in some ways has been my effort to show how the theses are jointly consistent. Indeed, what generates the paradox often has more to do with perceived inconsistencies among the theses. Recall that they are as follows:

1. *The HOT Thesis*: A version of the HOT theory is true (and thus a version of reductive representationalism is true).
2. *The Hard Thesis*: The so-called hard problem of consciousness, that is, the problem of explaining exactly how or why subjective experiences are produced at all from brain activity (or from any combination of unconscious mental activity), can be solved.
3. *The Conceptualism Thesis*: Conceptualism is true, that is, all conscious experience is structured by concepts possessed by the subject.
4. *The Acquisition Thesis*: The vast majority of concepts are acquired, though there are a core group of innate concepts.
5. *The Infants Thesis*: Infants have conscious mental states.
6. *The Animals Thesis*: Most animals have conscious mental states.
7. *The HOT-Brain Thesis*: There is a plausible account of how my version of HOT theory might be realized in the brain and can lead to an informative neurophysiological research agenda.

In chapters 1 through 5, I argued for the HOT and Hard Theses. I defended the Conceptualism Thesis mainly in chapter 6, but also in chapters 7 and 8. I defended the Acquisition and Infants Theses in chapter 7, and I argued for the Animals Thesis in chapter 8. Finally, in this chapter, the HOT-Brain Thesis took center stage. It would be silly to pretend that I have answered all of the questions on these very difficult and complex matters.

There have been, no doubt, occasions when I have had to navigate through some difficult waters, trying to keep in mind how subtleties with regard to one thesis might affect my defense of another thesis. However, at the least, I hope the reader agrees that I have made a philosophically cogent and empirically informed case for solving the Consciousness Paradox. It seems to me that there is good reason to be optimistic.

Notes

1 Introduction

1. A small sample of important single-author books includes Churchland 1986; Dennett 1991; Flanagan 1992; Crick 1994; Chalmers 1996; Lycan 1996; Baars 1997; Carruthers 2000; Tye 2000; Levine 2001; Koch 2004; and Zahavi 2005. Among the best anthologies on consciousness are Metzinger 1995 and Block, Flanagan, and Güzeldere 1997. Moreover, annual conferences such as "Toward a Science of Consciousness" and the "Association for the Scientific Study of Consciousness," as well as the journals *Philosophical Psychology*, *Journal of Consciousness Studies*, *Consciousness and Cognition*, and *Psyche*, have offered quality forums for disseminating work in the field. See also Gennaro 2005a.

2. Beginning with Rosenthal 1986. See especially his 2005 collection of essays.

3. For those interested, I also critically examined Sartre's theory of consciousness in light of HOT theory (Gennaro 2002) and argued that HOT theory can help us to understand Leibniz's theory of consciousness and self-consciousness (Gennaro 1999).

4. E.g., Fodor 1998; Cowie 1999; Prinz 2002; Murphy 2002; Bermúdez 2003; Gellman 2003; Nichols and Stich 2003; Rakison and Oakes 2003; Goldman 2006; Hurley and Nudds 2006.

5. E.g., Metzinger 2000; Baars, Banks, and Newman 2003; Cleeremans 2003; Gunther 2003; Gendler and Hawthorne 2006; Zelazo, Moscovitch, and Thompson 2007; Velmans and Schneider 2007; Bayne, Cleeremans, and Wilken 2009.

6. In a series of papers culminating in Kriegel 2009a.

7. For much more on the disputed issues involved, see the essays in Wright 2008.

8. Kriegel (2009a, 45) credits Levine (2001, 6–7) for previously making "essentially the same" distinction, but it is not at all clear to me that Levine agrees with Kriegel about the nature of the "for-me" or "subjective" component.

9. For much more on these definitional matters, see "Defining Consciousness," special issue, *Journal of Consciousness Studies* 16, no. 5 (2009); as well as De Quincey 2006 and the peer commentary that follows. For a nice discussion of some further subtleties, see Janzen 2008, chaps. 2 and 3; and Kriegel 2009a, chap. 2. For more on the terminological mess *among* several prominent higher-order theorists, see Byrne 2004, which clarifies and compares several additional distinctions, such as thin versus thick phenomenality (for Rosenthal), worldly versus experiential subjectivity (for Carruthers), and lower-order versus higher-order what-it's-like (for Lycan).

2 In Defense of the HOT Thesis

1. Higher-order representationalism was also arguably anticipated by Leibniz (Gennaro 1999) and Kant (Gennaro 1996). This idea has been revived over the past few decades by a number of philosophers, including Armstrong 1968, 1981; Lycan 1996, 2001a; and Rosenthal 1986, 1997, 2005.

2. See Husserl 1913/1931; Sartre 1956; Smith 1986, 2004.

3. The literature contains numerous excellent summary articles of these kinds of arguments, e.g., Gennaro 2005a; Kriegel 2007a; Levine 2007; Rowlands 2007; Kirk 2009; Stoljar 2009. See also Block and Stalnaker 1999; Hill and McLaughlin 1998; Perry 2001; and Kirk 2005. Some authors, for example, argue that some things *seem* possible but really aren't. Much of the debate centers on various alleged similarities or dissimilarities between the mind–brain and water–H_2O cases (or other scientific identities). Indeed, the issue of the exact relationship between "conceivability" and "possibility" is the subject of an important anthology (Gendler and Hawthorne 2002). See also Shear (1997) for specific responses to the hard problem and Chalmers's counterreplies.

In response to McGinn, for example, one might first wonder why we cannot *combine* the two perspectives in certain experimental contexts. Both first-person and third-person scientific data about the brain and consciousness can be acquired and used to solve the hard problem. Even if a single person cannot grasp consciousness from both perspectives *at the same time*, why can't a plausible physicalistic theory emerge from such a combined approach? Second, it may be that McGinn expects too much, namely, grasping some "causal link" between the brain and consciousness. After all, if conscious mental states are ultimately *identical* to brain states, then there may just be a "brute fact" that really does not need any further explaining. McGinn's argument may even presuppose some form of dualism to the extent that brain states are said to "cause" or "give rise to" consciousness, as opposed to using the language of identity.

Much the same goes for Frank Jackson's well-known (1982) "knowledge argument" against materialism. Jackson asks us to imagine a future where a person, Mary, is kept in a black-and-white room from birth, during which time she becomes a brilliant neuroscientist and an expert on color perception. Mary never sees red, for

example, but she learns all the physical facts and everything neurophysiologically about human color vision. Eventually she is released from the room and sees red for the first time. Jackson argues that it is clear that Mary comes to learn something new, namely, what it is like to experience red. This is a new piece of knowledge, and hence she must have come to know some nonphysical fact (since, by hypothesis, she already knew all the physical facts). Thus not all knowledge about the conscious mind is physical knowledge. One materialist reply is that Mary does not learn a new fact when seeing red for the first time, but rather learns the same fact in a *different way*. There is only the one physical fact about color vision, but there are two ways to come to know it: either by employing neurophysiological concepts or by actually undergoing the relevant experience and so by employing phenomenal concepts. For a thorough airing of the key issues, see Horgan 1984; Van Gulick 1985, 1993; Ludlow, Nagasawa, and Stoljar 2004; and Alter 2007. It is noteworthy that Jackson (2004) himself no longer takes the argument to refute materialism.

4. For more on phenomenal concepts, see Papineau 2002 and Carruthers 2005, chaps. 2 and 5. For much more on the *arguments* in question, see Carruthers and Veillet 2007; Diaz-Leon 2008; and the essays in Alter and Walter 2007. Tye (2000) was a believer in phenomenal concepts but has recently changed his mind on the issue for reasons I will not articulate here (Tye 2009b).

5. One might even develop an alternative HO theory of consciousness, a "quotational" HO theory, which attempts to use such concepts to explain conscious states (Picciuto 2011). For my own part, however, I prefer not to rely so heavily on the existence of phenomenal concepts.

6. For more formal versions of Searle's argument and numerous replies, see, e.g., Van Gulick 1995a, 1995b; Kriegel 2003a; Graham, Horgan, and Tienson 2007; Shani 2007, 2008.

7. I distinguish thirteen such interpretations in Gennaro 1995.

8. To be fair, both Terry Horgan and Charles Siewart have acknowledged to me in conversation that PI is indeed consistent with reductionism. For further discussion of this issue by another HO theorist, see Lycan 2008.

9. I use the phrase "aware of" instead of Rosenthal's "conscious of" mainly to avoid jargon and potential confusion as well as any appearance of circularity or regress. Rosenthal, of course, has in mind intransitive state consciousness being explained in terms of transitive consciousness.

10. Kriegel 2009a. For two actual *arguments* for the TP, see Janzen 2008, 69–84.

11. Thus I obviously disagree with Siewart (1998, 194–202) that something like the TP results from what he calls the "conscious-of trap" purely based on misleading language or an unjustifiable interpretation. It also seems to me that Siewart sometimes conflates unconscious HOTs with reflection or inner attention (i.e., conscious HOTs).

12. See also Millikan 1984 and Papineau 1987. Once again, there are some excellent overview articles on causal theories, such as Rupert 2008 and Adams and Aizawa 2010. It may well be that something closer to conceptual role semantics (CRS) is more plausible as an account of mental content for some other kinds of concepts, such as nonexistent objects and logical relations. According to CRS, the meaning of propositional content is determined by the role it plays in a person's language or cognitive system. The content of a representation is at least partly determined by the inferential connections that it bears to other representations. Overall, however, I take this view to be far less plausible for empirical concepts.

13. I revisit this overall theme of mental content and concept acquisition later, especially in chaps. 6 and 7.

14. I may be taking some liberties here in my characterization of "Fregean content" since the expression is sometimes used instead to refer to a condition on extensions rather than a psychological mode (or "manner") of presentation (see Chalmers 2004, 171–173). Nonetheless, what Frege himself meant by "mode of presentation" and how it is related to "sense" is not always clear either. I am most interested in the way that our concepts determine how one experiences outer objects and properties.

15. Narrow content has also seen resurgence in recent years among both reductionists and nonreductionists (see Rey 1998; Segal 2000; Horgan and Tienson 2002; Prinz 2002; Chalmers 2003, 2010; and Kriegel 2008). Indeed, as we have seen, one prominent HOT theorist explicitly defends narrow content (Carruthers 2000, 85–86, 105–113). Some of the more technical discussion revolves around so-called two-dimensional semantics, which recognizes two dimensions of the meaning or content of linguistic items. In this approach, expressions and their utterances are associated with two different sorts of semantic values that play different explanatory roles. Typically, one is associated with reference and ordinary truth conditions, while the other is associated with the way that reference and truth conditions depend on the external world. I will not pursue this theme further here.

16. It is also well known that allowing for narrow content helps to deflect the force behind so-called *inverted spectrum arguments* against wide content. Much the same goes for Block's (1990) well-known Inverted Earth argument, whose main target is wide content. I will not rehearse these arguments here but will say more about them in the next chapter. See Byrne 2010 for an excellent overview.

3 Assessing Three Close Rivals

1. But I return to it in chap. 5 in a somewhat different context.

2. Actually, the exact nature or even existence of nonconceptual content of experience is itself a highly debated and difficult issue in philosophy of mind. I address the topic at length in chap. 6.

3. See Lycan 2005 and Seager and Bourget 2007 for two nice overviews.

4. Rosenthal does not address inverted qualia at much length (2005, 149–152, 157–159, 223–226). He seems sympathetic to narrow content but is mostly concerned about whether or not such inversions could be detected from the third-person point of view. Once again, see Rey 1998 and Lycan 2001b for many of the pros and cons of phenomenal externalism. For another, more recent battle on these matters, see Pautz 2006 and Byrne and Tye 2006.

5. In chap. 6, I make a similar case for HOT theory over FOR with regard to our perception of ambiguous figures.

6. As we will see in the next chapter, this aspect of Carruthers's view is similar to my wide intrinsicality view (Gennaro 1996, 2006a).

7. Some of the points made hereafter can also be found in Rosenthal 2004, 24–30.

8. As Picciuto (2011) points out, Van Gulick seems to be equally susceptible to this objection. This is because the idea that there is "global availability" of some content to higher-order awareness seems to be a dispositional property.

9. Another interesting hybrid theory is offered by Lurz (2004), who calls his theory *same-order representationalism* (SOR). Very briefly, in contrast to both FOR and HOR, Lurz argues that a state C becomes conscious when another state M is immediately aware of the intentional *content* of C. I will not discuss SOR at length, but one problem I see is that even if it explains what makes the content of a state conscious, it does not explain what makes the *state* (or vehicle) itself conscious. This is a similar objection to the one raised earlier against FOR. It seems possible to have two mental states with the same content but in different modalities, such as *seeing* something flying overhead and *hearing* something flying overhead. But surely these conscious states have very different qualia although their intentional contents are the same. Similarly, being aware of intentional contents is not enough because the same content can be shared by different mental states, such as a *belief* that there is a cookie in the jar, a *perception* that there is a cookie in the jar, a *desire* for the cookie in the jar, *hoping* that there is a cookie in the jar, and so on. SOR cannot account for the different conscious states that result from a HOR directed at them. If Lurz really means for SOR *also* to include awareness of the states themselves, then it is unclear how it differs from HOR. Finally, most of Lurz's examples have to do with beliefs and language use, such as in responding to questions about one's beliefs. There are two problems here. One is that it is generally unclear what sense to make of first-order conscious beliefs, as I argued in the previous chapter. It seems to me that he is really addressing *introspecting* beliefs. The other is that by using sophisticated examples involving linguistic reports, Lurz seems to be restricting SOR to a narrow range of human consciousness.

10. See also Rosenthal 2004, 20–23. For another sustained critique of the "inner sense" model of introspection, see Shoemaker 1994.

11. For much more on Kant and HOT theory, see Gennaro 1996, chap. 3. I return to this theme in chap. 6 of this volume.

4 From HOT Theory to the Wide Intrinsicality View

1. There are, I believe, other significant advantages to something closer to the WIV. See Kriegel's discussion of the problems of immediacy and relationality (2006, secs. 4 and 5) and Van Gulick's (2004) criticism of standard HO theory's problematic handling of qualia.

2. Kriegel agrees that this is largely a terminological matter, but he still opts to restrict use of "higher-order" to theories that treat the HOT as a distinct state. Thus he often calls his view "same-order monitoring" (e.g., Kriegel 2006). However, during a session at the 2004 "Toward a Science of Consciousness" conference in Tucson, it also became clear that some hold stronger views on this matter. Andrew Brook urged me to jettison all use of "higher-order" in my theory, whereas Peter Carruthers thought that Kriegel had misnamed his theory. I agree more with Carruthers here, and Van Gulick also clearly has this preference; but, again, I take this mainly to be a terminological dispute. One problem, though, is that the converging similarities between all these positions might be lost.

3. By Andrew Brook and Robert Lurz, in conversations.

4. In retrospect, perhaps I should have chosen a more catchy name for my theory, but at this point, I hesitate to add to the abundance of acronyms and theory names already in the literature. Even just "WIT" (wide intrinsicality *theory*) would at least have been easier to *say*. Other, sexier possibilities are "intrinsic HOT theory" (IHOT) and the more provocative "1½ order theory of consciousness" or "split-level theory of consciousness." For what appears to be an intrinsic version of HOP theory, see Lormand (unpublished).

5. See, e.g., Rosenthal 2005, 10–14, 139–144, 168–171, 198–226.

6. For another important discussion of some of the themes is this section, see Matey 2006. Matey criticizes Rosenthal's HOT theory for holding what seems to be an inconsistent set of propositions, namely, (a) the lower-order state becomes conscious when a HOT is directed at it, (b) misrepresentations and targetless HOTs are possible, and (c) the lower-order state only become conscious when its content matches the HOT. Once again, although I agree with some of her criticisms of Rosenthal's HOT theory, it seems to me that she does not sufficiently allow for an alternative like the WIV. In addition, I reiterate that some of the discussion of targetless HOTs seems to have more to do with introspective consciousness. In some ways, then, much of this entire chapter can be seen as a detailed response to Matey's criticisms of standard HOT theory (especially later in section 4.5).

7. But see Gennaro 1996, 36–43, for one such attempt.

8. See, e.g., Rosenthal 1986, esp. 340–348; and Rosenthal 1997, 735–737. Intrinsic properties, as defined by Rosenthal, clearly need not be essential properties: my having dark hair is intrinsic to me but not essential. And conscious mental states can also have an informative and analyzable structure even if the HOTs are intrinsic to them (see Gennaro 1996, 21–30, for more on these and related points; cf. Schröder 2001, 33–34). Indeed, there is no logical connection at all between notions such as "intrinsicality," "extrinsicality," and "essentiality."

9. For a brief reply by Rosenthal on Block's "liver" version of the problem of the rock, see Rosenthal 2000c, 241.

10. I take the argument of this section also to be a response to Robinson 2004, 92–96, but I will not elaborate here. I also take the argument of the previous two sections to answer Mandik's 2009 anti-HOR "unicorn" argument. In short, at least one version of HOT theory, namely the WIV, can make sense of the notion that a mental state (MET) can represent a state (M) while both M and MET comprise a single conscious state. Much the same is true for Gois 2010, a paper that I became aware of shortly before this book went to press (see especially Gois 2010, 149–154). Of course, I do agree with the authors cited in this note to the extent that their criticisms of Rosenthal's theory are similar to some of my own. I elaborate further on these themes in section 4.5. Finally, the same applies to Block's (2011) critique of Rosenthal's HOT theory. Needless to say, I do not agree that the "higher-order approach" is itself "defunct."

11. See also Shear 1997, where Chalmers acknowledges that this problem has been recognized by many philosophers in the past under different names. More recently, Seager (1999) calls his version of the hard problem "the generation problem."

12. For example, Stubenberg cites Lycan's surprising statement that the "inner-sense account affords the best known solution I know to the problem of subjectivity and 'knowing what it's like'" (Lycan 1996, 15). See also Rosenthal 2002a, sec. IV. However, to be fair to both Lycan and Rosenthal, it is crucial to separate their explanations of sensory or qualitative properties *as they use these terms*, which are offered in a way so as to allow for these properties to be unconscious (see also e.g., Lycan 1996, 75–77; Rosenthal 2000c, 235–236), from their explanations of "knowing what it is like" to experience those properties from the first-person point of view. It is only for the latter "knowing what it is like" explanation that their respective versions of HO theory are invoked. For my own part, as I explained in chapter 1, I reserve the terms "sensory" and "qualitative" for the subjective first-person aspects of consciousness, but this is largely (though not entirely) a terminological difference. Part of the point of the rest of this section is to show that we can use HOT theory to explain consciousness and to address the hard problem in ways that go beyond what other HO theorists have said up to this point.

13. Though I would at least like to think that I did implicitly address the hard problem in Gennaro 1996, esp. chaps. 3 and 4.

14. I use the standard A/B reference system to Kant's *Critique of Pure Reason*, referring to his A and B editions respectively.

15. The issue is actually more complicated than I have portrayed it. Kant also speaks of the bridging faculty of "imagination" as performing a "threefold synthesis," at least in the A edition: "The apprehension of representation as modifications of the mind in intuition, their reproduction in imagination, and their recognition in a concept" (A97). Nonetheless I am not really concerned with this level of Kantian exegesis. Notice, however, that he does speak of "reproduction" and "recognition," which are major features of METs.

16. For more detail, however, see chaps. 6 and 7.

17. No doubt some will here be tempted to respond with a belief in some version of the so-called nonconceptual content in experience. I am firmly in the conceptualist camp, but a full response to this line of argument will have to wait until chapter 6. My main purpose here is to bring this Kantian-style version of the HOT theory to bear on the hard problem. Recall also that a HOT theorist argues that FOR cannot adequately or equally explain the difference between unconscious and conscious mental states. This is again why the concepts in question must be in the HOTs themselves.

18. But see Gennaro 1996, 54–68, for an initial attempt to address some of these questions. See esp. chaps. 6 and 7 in this book.

19. Once again, see, e.g., Papineau 1998; Block and Stalnaker 1999; Loar 1999; Yablo 1999; Carruthers 2000; Perry 2001; Botterell 2001; Kirk 2005, 2009.

20. I have already given advantages and motivations of HO theory over FO theory in previous chapters. My point here is not to argue for the HOT theory anew; it is mainly to address the Chalmers-style challenge that the denial of the HOT theory does not result in an explicit contradiction.

21. Witness Chalmers's remark that he is outlining "a two-dimensional intensional framework for handling a posteriori necessity" (1999, 435). For those unfamiliar with Chalmers's terminology here and for more details on the distinction between primary and secondary intensions, see Chalmers 1996, 52–71. For example, Chalmers explains that "the primary intension picks out a referent of a concept in a world when it is *considered as actual* . . . whereas the secondary intension picks out the referent of a concept in a world when it is *considered as counterfactual*, given that the actual world of the thinker is already fixed" (60). Perhaps most relevant here, however, is when Chalmers says that "we might as well think of the primary and secondary intensions as the a priori and a posteriori aspects of meaning, respectively" (62).

22. See, e.g., Kriegel 2006, n34; 2004a, n24.

23. It seems to me that, at least sometimes, Van Gulick's (2000, 2004, 2006) HOGS model is *arguably* committed to an unconscious part of a global conscious state. The

basic idea behind his "higher-order global states" theory is that "lower-order object states become conscious by being incorporated as components into the higher-order global states (HOGS). . . . The transformation from unconscious to conscious state is not a matter of merely directing a separate and distinct meta-state onto the lower-order state but of 'recruiting' it into the globally integrated state" (2004, 74–75). It is difficult to understand Van Gulick's frequent use of the expression "implicit meta-intentionality" or "reflexive self-awareness" of HOGS unless we (at least often) think of it as an unconscious part of a conscious whole. Much like my METs, which are implicitly (i.e., unconsciously) intrinsic to the structure of complex conscious states, such meta-intentional parts of HOGS might be taken to be unconscious. They are "implicit" and not "explicitly" contained in the structure of conscious experience. But Van Gulick normally speaks neutrally about whether or not the meta-intentional parts of his HOGS are conscious or unconscious, such as when he frequently says they are "built into" or "embedded in" the HOGS. This might lead one to construe his HOGS as similarly containing an unconscious metapsychological aspect. However, if Van Gulick means to suggest that such meta-intentionality is *consciously* part of one's outer-directed phenomenal experience, then I disagree for the same reasons I give in the next chapter. Some passages suggest such an interpretation; for example, "Experience is always the experience *of a self* and *of a world of objects*" (Van Gulick 2004, 81). And Van Gulick later explicitly says, "Nor is the HOGS model intended as a reductive theory" (2006, 37). Like Kriegel, Van Gulick is here clearly endorsing a nonreductive account of conscious states. Although I sympathize with the Kantian flavor of Van Gulick's account, this goes a bit too far, in my opinion. It is one thing to say that the HO state is "built into" or "presupposed in" the structure of conscious experience, but quite another to hold that it is itself a conscious part of a conscious state. I have much more to say on this topic in the next chapter.

24. See also Hossack 2003, 192, 200; Kriegel 2003b, 2004; D. Smith 2004, chap. 3.

25. I have in mind Kriegel (2005), Caston (2002, 777n54), and Ken Williford (e-mail communication, 2003); but recall also, on the other hand, that some accuse me of really holding some form of FO theory.

26. This point is echoed in Van Gulick's (2000, 2004) discussion of Chris Hill's "volume control hypothesis," whereby introspection is an active process in the sense that it often alters its lower-order mental object. I agree with Van Gulick and Hill here; however, they are clearly discussing *introspection* only. What is really needed is *also* the view that *unconscious* METs (or HOTs) affect the nature of their target states in cases of *first-order* conscious states.

27. See also Gennaro 1996, sec. 6.3, where I discuss Dennett's Chase and Sanborn example. Dennett describes two coffee tasters for Maxwell House. Mr. Chase reports that Maxwell House still tastes to him as it did when he first came to work there, but he no longer likes *that taste*. Mr. Sanborn also doesn't like the way it tastes to him now, but says it no longer tastes the same as it did when he started working there.

Unlike Dennett, I was not interested in showing that there are no qualia as such. Rather, I argued that we cannot separate out the "taste itself" or conscious sensation from the "attitude" or "judgment" toward it, the attitude or judgment being a MET and an intrinsic part of the conscious state.

28. Kriegel (2009a, 223n36) does indeed acknowledge this.

29. Also, as a proponent of narrow content, there seems to be the acknowledged advantage that this view does not have as much difficulty in explaining the causal efficacy of conscious states as it does for those who deny narrow content. Recall from chapter 2 that, in the externalist view, it was difficult to make sense of causal explanations as well as the notion that intrinsic duplicates have the same causal capacity to affect the world. See Kriegel 2009a, 140–142, for a nice discussion.

30. I have done so extensively elsewhere (Gennaro 1996, chaps. 3, 4, and 9).

31. Weisberg's subsequent article on misrepresentation (2011) does not really advance the debate much further. For example, he presents a false dilemma between accepting some kind of inexplicable intrinsic theory and accepting the Transitivity Principle. Weisberg also again misses the point that something more like the WIV can happily acknowledge that fallibility between HOT and M arises at the level of conscious HOTs or *introspection*. Indeed, he uses examples of misrepresentation that are best construed as introspective errors, such as when I mistake my anger at something at work as anger at the poor play of a football team. All should agree that we can misrepresent our introspective states; claiming otherwise is a straw-man argument against an intrinsic version of HOT theory. Finally, he mistakenly draws the analogy between, say, outer hallucinations (of pink elephants) and targetless (unconscious) HOTs. Once again, the proper analogy is between conscious first-order, or world-directed, states and conscious HOTs, or introspective states. See also Kidd (2011) for yet another discussion of infallibility and misrepresentation.

5 Against Self-Representationalism

1. Kriegel 2003b, 2003c, 2005, 2006, 2009a. See also the Kriegel and Williford 2006 anthology and the *Psyche* 2006 online symposium. I take what follows also to be decisive against Williford 2006.

2. Brentano's own view is actually much more elaborate than this. For example, Brentano (1874/1973, bk. 2, chap. 3) actually holds that *four* different aspects are included in every mental act, including a feeling toward itself. However, for my purposes, these problematic details are irrelevant and can safely be ignored.

3. It should be noted that Jean-Paul Sartre is also often cited as holding PSR. In Gennaro 2002, I argue that he held, or should have held, something closer to the WIV, though I did not explicitly address PSR in that paper. I will not address Sartre's view here.

4. This is therefore similar to Kriegel's SOMT2 (Kriegel 2006); but, as we shall see, I actually hold something closest to his SOMT10. Recall from chapter 4 that when I say that M and M* are "proper parts" of CMS, I basically mean that they are parts that are not identical with the whole of which they are parts. Thus, for example, M* cannot be part of itself; nor can the CMS be part of M*. It might be objected that my definition is circular, or at least nonreductionist, since the term "conscious" appears on each side of the biconditional. However, I do not think that this is case. The WIV is still reductionistic because an unconscious M* (= MET) is what makes an otherwise unconscious M conscious. The definition is also not circular because there is a crucial ambiguity in how the term "conscious" is used. The "conscious M" is meant to refer to the first-person subjective point of view, whereas speaking of the "whole CMS" is based more on third-person considerations. I chose to put M, instead of CMS, on the left side of the biconditional to more clearly express the differences among the three definitions (unlike Kriegel's SOMT10, which begins with the "whole" on the left side of the biconditional).

5. This is virtually the same as Kriegel's SOMT1 (in Kriegel 2006).

6. For some discussion of Sartre on this matter, see Gennaro 2002, 299–308.

7. David Woodruff Smith (e.g., 2004, chap. 3) is also a good example of someone concerned solely with the structure (or "form") of conscious states at the expense of offering any kind of explanation as such.

8. Once again, it may be useful to distinguish between "momentary focused introspection" and "deliberate introspection" (Gennaro 1996, 19–21). Momentary focused introspection is less sophisticated and only involves a brief conscious MET, whereas deliberate introspection involves the use of reason and a more sustained inner-directed conscious thinking over time. I mainly have momentary focused introspection in mind throughout this chapter.

9. For another, more sympathetic Brentanian attempt to allay such worries, see Textor 2006.

10. I should note here that Kriegel (2002b, 525) does briefly mention a possible Brentanian account of introspection. However, numerous questions remain regarding how faithful such an account is to Brentano and how it would really help to save PSR (hence I leave the question mark in fig. 5.1). Moreover, this and the previous objection to PSR are vividly illustrated in the writings of Hossack (2002, 2003). As an apparent supporter of PSR, Hossack frequently conflates first-order consciousness with introspection. He argues for an "identity thesis" defined as "each state of which one can be conscious is numerically identical with one's *introspective knowledge* of the occurrence of that very state" (2002, 163; italics mine). The problem is that Hossack often seems to have in mind inner-directed conscious focus when speaking of "introspective awareness" and "self-knowledge" (cf. Hossack 2003, 196). Indeed, he goes so far as to hold that the "Identity Thesis says that every experience and every

action is a conscious state, identical with knowledge of its own occurrence" (2002, 174). Thus Hossack faces the following dilemma: either such "self-knowledge" of a conscious state is outer directed, or it is inner directed. If it is outer directed, then he is not providing an analysis of M as part of an attempt to defend PSR; that is, M* would not be directed at M anyway. If it is inner directed (as it appears), then not only is he conflating M* with introspective awareness, but he clearly cannot justify identifying M* with conscious state M because M is an outer-directed state.

11. Moreover, Kriegel's subsequent discussion of "three positions" (2003b, 116ff) does *not* exhaust the possibilities. As he knows (Kriegel 2006), a fourth option is to treat M and M* as proper parts of a complex conscious state, even if we differ about whether or not M* is itself conscious.

12. D. Smith 2004, 78–79, 93. See also Thomasson 2000, 192, 196. Caston (2002, 792–793) also discusses the importance of the part–whole relationship in his analysis of Brentano and Aristotle. Perhaps most interesting is Caston's diagram (2002, 778) presumably representing PSR. But much like the WIV for outer-directed conscious states, we have an arrow going from one part of a complex conscious state (the "perceiving") to another part (the "seeing") divided by a broken line, in addition to the arrow representing the outer-directedness of the entire conscious state. Caston then speaks of *both* aspects as essential to any token perception. Some of the foregoing authors could mean something more like "property" by "part," but it is not clear to me how this helps us to understand the nature and structure of conscious states, let alone how it could explain what makes a mental state conscious.

13. Kriegel has indeed confirmed to me (in correspondence) that he does not currently hold PSR as he did at the time of writing Kriegel 2003b. However, he does still hold that M* is itself conscious, which is my main target in the remainder of this chapter. It is also what seems now to be the main difference between our theories.

14. This is not to say that we are always consciously aware of everything that we consciously attend to, as seems to be one lesson learned from cases of inattentional blindness (Mack and Rock 1998). I will explore the relationship between consciousness and attention further in the next chapter.

15. See Kriegel 2004b, 177–178, for a nice discussion of this point.

16. See Lycan and Ryder 2003 and W. Wright 2005; but also see Janzen 2005 for discussion of the case of the long-distance truck driver.

17. It is worth noting that some who are well known for emphasizing the importance of phenomenological data do not side with Kriegel on this point. See, e.g., the discussion in Zahavi 2004a and Siewart 1998, chap. 6. These authors, however, also reject any form of higher-order theory. See also Lyyra 2009 for further discussion on this matter.

18. For related analysis of a few other psychopathologies, see Gennaro 1996, 136–142.

19. Much of the impetus behind this subsection resulted from some extremely helpful comments and questions when I presented a paper on this topic at the 2006 "Toward a Science of Consciousness" conference in Tucson, Arizona. Special thanks to Pete Mandik, Bob Van Gulick, Josh Weisberg, and Uriah Kriegel.

20. Another case might be sitting on my back patio enjoying the weather when I am interrupted by one of my young children running toward the street. When I see that, I'd say that all my consciousness becomes outer directed.

6 In Defense of Conceptualism

1. It is worth noting, however, that there is a very interesting ongoing debate about whether or not Kant himself was a conceptualist. See Hanna (2005, 2008) versus Ginsborg (2006, 2008). I will not pursue the issue here, but for the record, I side with Ginsborg, who argues that Kant was a conceptualist. See also Gennaro 1996, esp. chaps. 3 and 4, for more on this Kantian line of thought.

2. I say this although some authors do seem to embrace something like STRONG-NC. At the least, some embrace the related "autonomy thesis," which says that "it is possible for a creature to be in states with nonconceptual content, even though that creature possesses *no* concepts at all" (Bermúdez 1998, 61; italics mine).

3. But for much more on this and related distinctions, see also Byrne 2005; Laurier 2004; Speaks 2005; Crowther 2006; Heck 2007.

4. There are also two other main sources of support for NC. The "continuity argument" suggests that NC is true because it is the best way to account for similar nonconceptual experiences in animals and infants. Some have also suggested that concept acquisition is possible only if there are nonconceptual contents in experience (Roskies 2008). I address these arguments in the next two chapters.

5. But this was part of my motivation for editing Gennaro 2007.

6. As is common practice among philosophers, I heretofore adopt the convention of using small capitals when using a word or expression to refer to the concept. Thus DOG refers to the *concept of dog*, whereas the usual word (dog) refers to the animals themselves.

7. I do not mean to suggest that this is entirely uncontroversial. See Briscoe 2008 for an excellent critical discussion of the two-visual-systems hypothesis, including evidence from the Titchener circles illusion.

8. It is also worth mentioning that there is some analogous evidence for two systems in hearing (Rauschecker 1998).

9. Peacocke (1992) develops an interesting positive account of nonconceptual perceptual content in terms of what he calls "scenario content." Scenario content is a way of representing the space around a perceiver's body with surfaces, solids, textures, lights, distances, and so on, which are consistent with the veridicality (or correctness) of the experience. According to Peacocke, such contents need not be built out of concepts possessed by the subject but are also presumably personal-level content. A conceptualist, however, might view the matter quite differently. One possibility would instead be to treat such content as subpersonal spatial information involving one's body and its movements. Another possibility would be to allow that something like scenario content is deployed in conscious experience, but it involves combinations of very basic (even innate) concepts, such as SPACE, UP, DOWN, OBJECT, NUMBER, and so on.

10. By "determines" I mean not "causes" but "explains" or "accounts for."

11. This goes back to my 1996 discussion of Kant, which came about independently of McDowell's treatment.

12. See MacPherson 2006, sec. 7, for a response to Peacocke on the square/diamond example.

13. I think much the same argument can be used in response to a similar argument in Nickel 2007), but I won't elaborate here. He uses the example of a set of nine tiles or squares in the pattern of three rows of three. He points to the shift that occurs when one sees them phenomenologically grouped in two different ways, with different sets of tiles becoming more prominent. Once again, however, Nickel's target is FOR, not HOT theory.

14. Some version of this argument is presented by many authors, e.g., Heck 2000; Speaks 2005; Tye 2006b.

15. For more discussion, see Byrne 1997, 120; and Tye 2006b, 516–517.

16. This is related to a point made by Noë (2004, 48–49) in rejecting what he calls the "snapshot conception" of experience. For much more on this general theme, see Noë 2002, which is a special issue of the *Journal of Consciousness Studies*, asking the question "Is the visual world a 'grand illusion'"?

17. For much more on this ongoing debate, see, e.g., the special issue of *Psyche* 14 (2008); esp. Cavanna and Nani 2008 and Bartolomeo 2008.

18. I thank Joseph Gennaro for calling this difference to my attention.

19. I thank Deidra Gennaro for a helpful discussion of this example.

20. One might also wonder about whether or not so-called pure conscious events (PCEs) falsify CON or even HOT theory. PCEs are mystical states of mind that are allegedly empty of all experiential and conceptual content. Gunther (2003, 1–2) also refers to the Eastern tradition as sympathetic to NC. PCEs appear to be strongly at

odds with CON and HOT theory. After all, if PCEs are devoid of conceptual content, then they obviously cannot be conceptual through and through, as the conceptualist or HOT theorist claims. I won't address this theme in this book, but see Gennaro 2008b for a more direct and detailed response.

7 Concept Acquisition and Infant Consciousness

1. See Griffiths and Machery 2008 and Khalidi 2009 for another interesting debate on scientific value of innateness as a folk psychological notion.

2. Once again, Fodor (1998) seems to reject this extreme view. See Cowie 1999, pt. 2, for much more discussion on this point. I'll set aside these details, since my main focus is on the relationship between concepts and consciousness.

3. For the interested reader, Kant presents his categories, or the "concepts of the understanding," as four groups of three: Quantity (UNITY, PLURALITY, TOTALITY), Quality (REALITY, NEGATION, LIMITATION), Relation (SUBSTANCE, CAUSE, COMMUNITY), and Modality (POSSIBILITY, EXISTENCE, NECESSITY).

4. For a more detailed review, see, e.g., Murphy 2002, 272–283; Rakison and Oakes 2003, 14–18. For a defense of the methods used, see Carey 2009, 105-111.

5. See also Kinzler and Spelke 2007 for a concise overview of some of the relevant literature.

6. Similar results from other experiments on infants abound in the literature (Baillargeon 1987; Spelke 1990; Spelke and Van de Walle 1993; Needham 2001). See also the famous "drawbridge experiment" in Baillargeon, Spelke, and Wasserman 1985, which uses the violation-of-expectation method. See also Carey 2009, chaps. 2 and 3, for additional discussion.

7. Xu and Carey 1996; Xu, Carey and Welch 1999; but see also Needham and Baillargeon 2000; Xu and Carey 2000; and Xu, Carey, and Quint 2004.

8. For some interesting related discussion, see the 2008 *Philosophical Psychology* symposium, especially the Keil and Mandler essays, on what extent a congenitally blind person could still have a notion of SPACE via nonvisual sensory input.

9. For excellent discussions of the relationship between time and perceptual experience, see Dainton 2000; Noë 2006; Le Poidevin 2009; and Hoerl 2009. For a more Kantian-style argument, see Gennaro 1992; 1996, chap. 9.

10. For further discussion, see also Carey 2009, chap. 6.

11. For much more along these lines, see Bloom 2000, which defends an essentialist theory of concept learning and representation. See also Bloom 2001, which includes a target article and open peer commentary. Gelman (2003) argues at book length for essentialism, namely, that it is an early cognitive bias.

12. See, e.g., Mandler's 1999 reply to Madole and Oakes 1999; McDonough and Mandler 1998; and Mandler 2004, 182–188.

13. For other useful summaries and reviews, see Seger 1994; Frensch and Runger 2003; Reber 1993; and Stadler and Frensch 1998.

14. For more on this issue, see Goldstone and Barsalou 1998; Madole and Oakes 1999; Carey 2000; and Murphy 2002, 295–302.

15. Roskies also says that she favors the "state" (as opposed to "content") view of conceptualism (2008, 650–651), but I will ignore this issue here. It is not clear that much depends on this. See sec. 6.1 in this volume.

16. But see Clark 2006 as well as Campbell's 2006a and 2006b replies to both Clark and Matthen. See also Raftopoulos 2009b, 343–350, on this point.

17. Roskies simply dismisses, for example, Raftopoulos and Müller's 2006 critique of Campbell without argument (Roskies 2010, 133n13).

18. Some authors do speak of "pre-reflective self-consciousness" (e.g., Legrand 2007b), but they sometimes seem to mean something closer to the nonreductive self-representationalist account rejected in chapter 5.

19. There are numerous abnormal adult pathological conditions with respect to self-consciousness and I-concepts. For example, schizophrenia is perhaps the most discussed along these lines. See Frith 1992; Stephens and Graham 2000; Zahavi 2000; Jeannerod 2004 (83–84); Gallagher 2004; and Stephens and Graham 2007.

20. For much more on this paradigm and results, see Rovee-Collier, Hayne, and Colombo 2001, chap. 6. See also chap. 8 for additional evidence using mobiles. See Hayne 2007 and Bauer et al. 2007 for additional compelling results.

21. See also Mandler's discussion of GOAL and evidence for very early infant attribution (2004, 102–108; 2008, 216–217). It should be noted here also that Rakison (2007) critically examines the developmental literature in the areas of mathematics, categorization, and induction in order to determine whether infants possess concepts that allow them explicitly (i.e., consciously) to reason and make inferences about objects and events in the world. He argues against the idea that infants have conscious access to such background knowledge and then speculates about the relationship between language development and consciousness. However, he uses the terminology of "conscious access" to mental states and background knowledge, which is more like what we have termed introspection. As we have seen, introspection is not needed for conscious states generally, even according to HOT theory. Moreover, Rakison does acknowledge that infants are phenomenally conscious and capable of, say, feeling pain.

22. For more evidence along these lines, see Surian, Caldi, and Sperber 2007.

8 Animal Consciousness

1. For much more along these lines, especially on comparative neuropsychological evidence, see Baars 2005; Griffin and Speck 2004; Edelman, Baars, and Seth 2005; Beshkar 2008; and Edelman and Seth 2009. For example, numerous animals, including some nonmammals, have some form of thalamocortical structure that is sometimes held to be a locus of conscious experience in humans (Baars 2005; Edelman, Baars, and Seth 2005).

2. See Allen 2004 and Shriver 2006 for much more on animal pain, including recent empirical research designed to challenge skeptics.

3. This chapter is also therefore designed to respond to Tye's view, as well as to Allen-Hermanson 2008. It is also perhaps terminologically confusing for Tye to talk about phenomenally conscious pains without suffering, but I leave that aside here. Allen-Hermanson (2008), however, also argues that FO theorists face some difficulties of their own in accounting for animal consciousness.

4. There are other reasons for the belief in animal consciousness, such as arguments based on an inference to the best explanation or an argument from analogy between humans and some animals. For some relevant overall reviews and summaries, see Allen 2010; Andrews 2008; Beshkar 2008; Lurz 2009a. For additional evidence that most animals have conscious *emotions*, such as fear, grief, and hope, see R. Roberts 2009.

5. Once again, theory-theory is usually contrasted with "simulation theory." In fact, however, many theorists hold some form of hybrid theory (Carruthers and Smith 1996; Nichols and Stich 2003; Goldman 2006). The term "theory of mind" goes back to Premack and Woodruff (1978), who used it with reference to whether chimpanzees are able to attribute beliefs and desires to others to predict and explain their behavior.

6. Tulving even acknowledges (2005, 26) that KC clearly has metacognitive abilities, and believes that he would pass various tests that measure the ability to understand other minds (or "theory of mind" abilities).

7. See also Zentall 2005 and DeGrazia 2009 for further evidence both for EMs and for anticipation of future events in various animals. See Raby and Clayton 2009 for an interesting related discussion. Finally, see Shea and Heyes 2010 for an argument that metamemory is evidence of animal consciousness.

8. For much more on all the controversy, see Suddendorf and Corballis 2007, which includes peer commentary and authors' response. In addition, see Raby and Clayton 2009, 319–322, for a more recent reply to Suddendorf and Corballis. In short, it often seems that those who doubt animal episodic memory utilize unnecessarily high requirements for having episodic memory or the ability to "mentally time travel" into

the past (or the future). Suddendorf and Corballis also seem not to allow properly for different *degrees* of mental time travel across different species.

9. I have intentionally avoided discussing the well-known mirror recognition test for various reasons (Gallup 1970; Keenan, Gallup, and Falk 2003; DeGrazia 2009). My own view is that while these results are interesting, it is not the best test for determining a clear type of self-awareness. However, it seems clear that even those who fail the test at least seem to be able to distinguish their bodies from the mirror itself. Thus a level-1 self-concept is still present.

10. Cf. Proust's distinction between what she calls "metarepresentation" and "metacognition" (2006, 260; cf. Proust 2009). However, this is different from the conscious and unconscious HOT distinction partly because, according to Proust, mental concepts only appear at the metarepresentation level.

11. Earlier similar results using a competitive paradigm are reported for chimps in Hare et al. 2000; Hare, Call, and Tomasello 2001; Tomasello, Call, and Hare 2003; Hare and Tomasello 2004; see also Tomasello and Call 2006. For example, subordinate chimps selectively tried to obtain food that dominant individuals could not see. See also Emery and Clayton 2009 for nice review and discussion of the themes addressed in this section.

12. For further defense of the view that self-attribution of mental states (metacognition) is prior to our capacity to attribute mental states to others (mindreading), see Goldman 2006. A more modest view, offered by Nichols and Stich (2003), is that the two capacities are independent and dissociable. Once again, however, most authors in this area seem really to embrace some kind of hybrid theory. Carruthers (2009b) argues at length that mindreading is actually prior to metacognition. For many of the reasons offered in both this and the previous chapter, I am not convinced that the evidence supports his view better, say, than Nichols and Stich's position. Recall that the two main opposing views are simulation theory (ST) and theory-theory (TT). ST holds that mindreading involves the ability to imaginatively take the perspective of another. TT holds that metacognition results from one's "theory of mind" being directed at oneself. So which of the three views is closest to the truth? I am frankly not at all sure that we have enough evidence to decide, but, I think it is safe to say that it is premature to suppose that mindreading is *prior to* metacognition, as Carruthers thinks. This would preclude the possibility of first-person metacognition dissociated from mindreading, for which there is significant evidence. For example, it is doubtful that autistic people have an *equal* impairment of mindreading and metacognitive abilities. The evidence seems to suggest that mindreading is lacking to a more serious degree than is metacognition generally. Indeed, in some cases, autistic people seem able to do surprisingly well on metacognitive tasks. Thus, I am most sympathetic with Nichols and Stich's analysis.

13. The case of deception is related to the false-belief task often used in research on infants and young children, as was discussed in the last chapter. In this case,

the subject is being asked to *recognize* a false belief in another instead of *cause* it. An object might be moved to another location while subject A is in the room but subject B is out of the room. When B returns, A might be asked where B will look for the object to determine whether or not A can successfully contrast its own belief to B's belief about where the object will be. Infants do not perform well on these tasks until at least age three. But, of course, they are often asked to *verbalize* their attitudes toward another's beliefs or perceptions, and surely that indicates a *conscious* attitude toward another's mental state. So the same point applies here: HOT theory allows for the presence of conscious states in the absence of (either self-attributing or other-attributing) *conscious* HOTs. Moreover, there at least seems to be something more sophisticated involved in grasping the concept of a mental state *mis*representing the world from merely representing the world. To the extent that an infant might have a primitive grasp of merely representing the world without misrepresenting the world, it seems wise to use caution in reading too much into these results. Andrews (2005) also interestingly argues that Povinelli and Vonk's (2004) critique of the food competition paradigm for chimps should logically lead them to the surprising, and perhaps absurd, conclusion that even when children *pass* the false-belief task, there is still *no* theory of mind for those children. A nonmentalistic story *could* still be given for such children.

14. For more on this line of argument and varieties of the generality constraint, see Carruthers 2009a. He argues that there are good reasons to suppose that animals (even invertebrates) can adhere to a weaker (causal) generality constraint, according to which genuine concepts must be recombinable with *some* others. Humans, on the other hand, are more sophisticated thinkers who are able to think more creatively than animals and can use language to do so.

15. Much as with infants, there is also the issue of what *kind* of mental state is being attributed. Mental-state attribution may not be an all-or-nothing affair. For example, there is reason to think, at least for humans, that having or attributing *beliefs* is a somewhat more sophisticated capacity than having other kinds of mental states, such a desire, perception, and pains (see Ridge 2001, 327, for some discussion.). Thus if an animal or infant fails one experiment testing belief attribution, it may well still be able to attribute desires or perceptions, not to mention merely *have* pains. This is perhaps similar to the distinction made by Bermúdez (2009) between perceptual mind reading and propositional-attitude mind reading.

16. Contra Carruthers (2008), who cites well-known problems with infallibility (e.g., from Nisbett and Wilson 1977) as one reason to opt for the opposing view. This is really not the issue; rather, the issue is whether or not it is typically *harder* to know about another's mental state. In other words, aren't we even *more often* wrong when attributing mental states to others? Also, Nisbett and Wilson were largely concerned with errors in *reasoning about* our mental states, not our awareness of them in the first place.

17. Sober's Canon is basically that faculty H is higher than faculty L if H entails L, but not conversely. I will not discuss it here. For some criticisms, see Allen-Hermanson 2005, 617–622; and Montminy 2005, 402–406.

18. For much more on this line of argument regarding primate mind reading, especially against Povinelli and Vonk (2006), see Fitzpatrick 2009. See also Penn and Povinelli 2007, along with Lurz's 2009c reply and his attempt to resolve the overall methodological problem of designing a food competition protocol aimed at further distinguishing between mind-reading and behavior-reading interpretations (cf. Hare et al. 2000; Hare, Call, and Tomasello 2001; Povinelli and Vonk 2004). See also Lurz 2011.

19. See also J. Smith 2005, 257–60, for a forceful and similar line of argument.

20. I will not go into detail here on comparative animal neurophysiology, but for much more along these lines, again see Baars 2005; Edelman, Baars, and Seth 2005; and Beshkar 2008. Beshkar (2008) also brings together a plethora of supporting evidence regarding animal tool use, communication, problem solving, and deceptive behavior, especially for various mammals, birds, spiders, and bees.

21. See Lurz 2007 for a similar but not identical reply to Bermúdez.

22. For a book-length study of empathy, including discussion of theory of mind and autism, see Stueber 2006.

23. For more discussion on this point, see Zahavi 2005, 192–196, 215–222.

9 Into the Brain

1. For much more on brain structure and on the function of the different brain areas, see Baars and Gage 2010.

2. For more, see Blackmore 2004, 228–229.

3. See Block 2007, 495–498, for further evidence. One might also consider dreams in this context. On what appears to be the reasonable assumption that at least some dreams are conscious, there is little, if any, PFC activity during those episodes (see Revonsuo 2006, chaps. 3 and 4). Moreover, it may be oversimplistic to refer to the PFC *as a whole*, since there are many subparts of the PFC, including the dorsolateral PFC (dlPFC) which is often the most specific area of interest.

4. On this general theme, see also Seth, Baars, and Edelman 2005; and Edelman, Baars, and Seth 2005. See Tononi 2004 for a related view, but one that emphasizes the role of informational complexity and integration in conscious states.

5. I should add here that there is also the *indirect* evidence mentioned earlier. That is, if one wishes to show that HOT theory is consistent with the available evidence, then it is necessary to look elsewhere in the brain.

6. I have not said much about the third conjunct, the claim that there is little or no PFC activity in infants and most animals. But even this is not as simple as it might seem. The PFC incorporates many subareas, some of which can be found in many animals. In previous work, I made more modest attempts to show that even if HOTs require the presence of some *cortical* structures, most animals do indeed have cortices or structures homologous to the human cortex. Many mammals even have a *neo*cortex. There is also a neocortex in infants and late fetuses. Many nonmammal vertebrates, such as reptiles and birds, also have some kind of cortex (see Gennaro 1996, 91–95).

7. For much more on this and other empirical and neurobiological theories of consciousness, see Kouider 2009; and Revonsuo 2010, chap. 11.

8. Beeckmans (2007) challenges HOT theory's neurological plausibility by pointing out that there is no evidence that, for example, conceptually detailed chromatic information (or conceptual short-term memory) is represented in the frontal and prefrontal lobes. As we have seen earlier in this chapter, however, HOT theory need not be committed to HOTs (or at least unconscious HOTs) being located in the PFC. Furthermore, as Beeckmans recognizes, a HOT theorist might also counter that the level of "detail" or "richness" in our visual experience is overstated, as was argued in chapter 6. To be fair, however, Beeckmans does entertain the possibility that a HOT theorist might invoke "ensemble concepts" to explain away the sense of chromatic richness (2007, 106–108). Ensemble concepts function very much like applied coarse-grained concepts and are often plural concepts, such as GREENS, LEAVES, NUMEROUS EDGES AT DIFFERENT ANGLES, when applied to a perception of a tree.

9. Inspired by Jackendoff (1987), Prinz (2000, 2007) has argued for what he calls the "attended intermediate-level representation" (AIR) theory of consciousness or the "intermediate level" theory of consciousness. Prinz recognizes that his theory may have some affinity with HOT theory, but in fact rejects HOT theory for a number of reasons that I won't articulate here (Prinz 2000, 255). Perhaps most relevant here, however, is that Prinz explicitly states that the "intermediate level" is necessary (but not sufficient) for having conscious states (2007, 257). The question then arises as to what else is needed for sufficiency. I suggest that something like the higher-order thought (HOT) theory of consciousness might therefore be a useful complement to AIR theory (which has been humorously dubbed the "HOT AIR" theory by Prinz). On the cognitive level, we might think of an unconscious HOT as similar to the kind of higher-level top-down attention described by Prinz. On my view at least, this would still not imply that the neural level realization of conscious states would therefore be too distributed or "high," including the PFC. Prinz seems to agree that conscious states do not require PFC activity. He thinks, for example, that attention is based in posterior areas. Thus, even though HOT theory demands that conscious states are distributed to some degree, a more moderate global view is preferable, especially with respect to first-order conscious states. Prinz seems open to the idea that HOT theory and AIR theory might at least be integrated in some way. For more on Prinz's

view and its relation to other theories of consciousness, see the commentary that follows Prinz 2000 and the related discussion at http://onthehuman.org/2010/11/does-consciousness-outstrip-sensation, including Prinz's replies.

10. Indeed, Bermúdez (2007b, 58–62) nicely explains how this could work in conjunction with a connectionist approach (cf. Munkata et al. 1997; Munkata and McClelland 2003). What I consider to be a concept, however, he may consider nonconceptual.

11. For a summary of the issues and for additional references, see Waskan 2010, sec. 6; and Garson 2010, secs. 5–8.

12. For much more of an overview on these issues, see Blackmore 2004, chaps. 7, 8, 17; Brook and Raymont 2010; and Revonsuo 2006, pt. 4.

13. Numerous interesting illustrations abound in Cleeremans 2003. For example, interesting effects are generated in the Dalmatian dog experiment, whereby neural oscillations differ as the subject experiences the coherent percept of a dog instead of mere meaningless black dots against a white background (156–158). In other cases, objects (such as a triangle) "pop out" to conscious experience owing to the unique feature of closure, whereas no such pop-out phenomenon occurs when there is a different conjunction of the same three lines (e.g., forming an arrow) against a similar background of lines (101–102).

14. On the other hand, several authors in Cleeremans 2003 (such as Hurley, Cotterill, and Valera and Thompson) take a more radical and somewhat skeptical position on finding NCCs and solving the binding problem. In particular, these authors raise serious questions as to whether we can solve the binding problem (and the search for NCCs) by looking exclusively at the brain. Consciousness, they say, is more of an entire bodily and motor activity that does not merely take place within the skull. This is the so-called sensorimotor or enactive theory of consciousness. Although brain activity may be necessary for consciousness (including the unity of consciousness), the idea is that we must resist the natural tendency to locate consciousness entirely within the brain. Consciousness is instead a capacity of the whole organism. We need to "go beyond the notion of a skull-centered correlate of consciousness to consider the multifarious ways in which brain processes are part of organismic cycles that generate the somatic, environmental, and social dimensions of our experience" (Varela and Thompson 2003, 282). If these authors are correct, then the search for necessary conditions of binding and consciousness *that are jointly sufficient* is doomed to failure because they do not account for other essential contributions of an animal's body. Their fundamental point is that "no neural process *per se* can be 'the place where consciousness happens' because conscious experience occurs only at the level of the whole embodied and situated agent. Neurons and [even] neural assemblies are not conscious subjects; persons and animals are" (Varela and Thompson 2003, 281). This is also what Hurley calls "the insight of *vehicle externalism*" (2003, 81). Thus, according to Hurley, the mechanisms of unity extend beyond the brain to the entire

body and through motor feedback. Moreover, any puzzle about partial unity (e.g., in split-brain cases) disappears when one recognizes that perception depends on action.

I am frankly not sympathetic to, and often puzzled by, this approach for a number of reasons, which would take me too far afield. For one thing, I am still not sure what it means to say things like "conscious experiences occur at the level of the entire organism" or that conscious mental states occur (partly) outside the skull. To be sure, the *content* and *causal interaction* of conscious states will frequently involve reference to bodily and motor elements, but that is still not to say that consciousness, or the vehicle of consciousness, is literally partly located outside the skull. This issue is also played out at length in a special issue of the *Journal of Consciousness Studies* 11, no. 1 (2004), under the title "Are There Neural Correlates of Consciousness?" There is a target article by Noë and Thompson, followed by commentaries and author response.

15. See, e.g., Sacks 1987; Ramachandran and Blakeslee 1998; Stephens and Graham 2000; Blackmore 2004, chap. 7; Revonsuo 2006, chaps. 12 and 13; Bayne 2008, 2010. On the issue of how problems with the unity of consciousness can manifest themselves temporarily in *everyday* memory failures, see Gennaro, Herrmann, and Sarapata 2006.

16. According to HOT theory, it certainly might be the case that two conscious states combine when there is introspection; that is, the higher-order conscious HOT combines with the conscious LO state to produce an all-encompassing conscious state. However, there is again little reason to suppose that there are two conscious experiences *in addition to* the combined conscious state.

17. The matter is even further complicated by puzzles about the nature of indexicals and purported analogies to linguistic self-reference, but I won't delve into these issues here. See Rosenthal 2003, 338–349, for some discussion.

18. For some discussion on Kant's "I think" and HOT theory, see Gennaro 1996, 48–54.

References

Adams, F., and K. Aizawa. 2010. Causal theories of mental content. In *The Stanford Encyclopedia of Philosophy* (spring 2010 edition), ed. Edward N. Zalta, http://plato.stanford.edu/archives/spr2010/entries/content-causal.

Aglioti, S., J. DeSouza, and M. Goodale. 1995. Size contrast illusions deceive the eye but not the hand. *Current Biology* 5:679–685.

Aguiar, A., and R. Baillargeon. 1999. 2.5-month-old infants' reasoning about when objects should and should not be occluded. *Cognitive Psychology* 39:116–157.

Alkire, M., A. Hudetz, and G. Tononi. 2008. Consciousness and anesthesia. *Science* 322:876–880.

Allen, C. 1999. Animal concepts revisited: The use of self-monitoring as an empirical approach. *Erkenntnis* 51:33–40.

Allen, C. 2004. Animal Pain. *Noûs* 38:617–643.

Allen, C. 2010. Animal consciousness. In *The Stanford Encyclopedia of Philosophy* (fall 2010 edition), ed. Edward N. Zalta, http://plato.stanford.edu/archives/fall2010/entries/consciousness-animal.

Allen, C., and M. Hauser. 1991. Concept attribution in nonhuman animals: Theoretical and methodological problems in ascribing complex mental processes. *Philosophy of Science* 58:221–240.

Allen-Hermanson, S. 2005. Morgan's canon revisited. *Philosophy of Science* 72: 608–631.

Allen-Hermanson, S. 2008. Insects and the problem of simple minds: Are bees natural zombies? *Journal of Philosophy* 105:389–415.

Alter, T. 2005. The knowledge argument against physicalism. In *Internet Encyclopedia of Philosophy*, http://www.iep.utm.edu/know-arg.

Alter, T. 2007. The knowledge argument. In *The Blackwell Companion to Consciousness*, ed. M. Velmans and S. Schneider. Malden, MA: Blackwell.

Alter, T., and S. Walter, eds. 2007. *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. New York: Oxford University Press.

Andrews, K. 2005. Chimpanzee theory of mind: Looking in all the wrong places. *Mind and Language* 20:521–536.

Andrews, K. 2008. Animal cognition. In *The Stanford Encyclopedia of Philosophy* (winter 2008 edition), ed. Edward N. Zalta, http://plato.stanford.edu/archives/win2008/entries/cognition-animal.

Armstrong, D. 1968. *A Materialist Theory of Mind*. London: Routledge & Kegan Paul.

Armstrong, D. 1981. What is consciousness? In D. Armstrong, *The Nature of Mind*. Ithaca, NY: Cornell University Press.

Ashby, F. G., and W. T. Maddox. 2005. Human category learning. *Annual Review of Psychology* 56:149–178.

Ashby, F. G., and B. Spiering. 2004. The neurobiology of category learning. *Behavioral and Cognitive Neuroscience Reviews* 3:101–113.

Baars, B. 1988. *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Baars, B. 1997. *In the Theater of Consciousness*. New York: Oxford University Press.

Baars, B. 2005. Subjective experience is probably not limited to humans: The evidence from neurobiology and behavior. *Consciousness and Cognition* 14:7–21.

Baars, B., W. Banks, and J. Newman, eds. 2003. *Essential Sources in the Scientific Study of Consciousness*. Cambridge, MA: MIT Press.

Baars, B., and N. Gage. 2010. *Cognition, Brain, and Consciousness: Introduction to Cognitive Neuroscience*, 2nd ed. Oxford: Elsevier.

Baillargeon, R. 1987. Object permanence in 3.5 and 4.5 month old infants. *Developmental Psychology* 23:655–664.

Baillargeon, R. 1994. How do infants learn about the physical world? *Current Directions in Psychological Science* 3:133–140.

Baillargeon, R. 2004. Infants' physical world. *Current Directions in Psychological Science* 13:89–94.

Baillargeon, R. 2008. Innate ideas revisited. *Perspectives on Psychological Science* 3:2–13.

Baillargeon, R. and J. DeVos. 1991. Object permanence in young infants: Further evidence. *Child Development* 62:1227–1246.

Baillargeon, R., E. Spelke, and S. Wasserman. 1985. Object permanence in 5-month-old infants. *Cognition* 20:191–208.

Balog, K. 1999. Conceivability, possibility, and the mind–body problem. *Philosophical Review* 108:497–528.

Baron-Cohen, S. 1995. *Mindblindness*. Cambridge, MA: MIT Press.

Bartolomeo, P. 2008. Varieties of attention and of consciousness: Evidence from neuropsychology. *Psyche* 14, http://theassc.org/files/assc/2255.pdf.

Bauer, P., T. DeBoer, and A. Lukowski. 2007. In the language of multiple memory systems: Defining and describing evelopments in long-term declarative memory. In *Short- and Long-Term Memory in Infancy and Early Childhood*, ed. L. Oakes and P. Bauer. New York: Oxford University Press.

Bayne, T. 2004. Self-consciousness and the unity of consciousness. *Monist* 87: 219–236.

Bayne, T. 2008. The unity of consciousness and the split-brain syndrome. *Journal of Philosophy* 105:277–300.

Bayne, T. 2009. Perception and the reach of phenomenal content. *Philosophical Quarterly* 59:385–404.

Bayne, T. 2010. *The Unity of Consciousness*. New York: Oxford University Press.

Bayne, T., and D. Chalmers. 2003. What is the unity of consciousness? In *The Unity of Consciousness: Binding, Integration, and Dissociation*, ed. A. Cleeremans. Oxford: Oxford University Press.

Bayne, T., A. Cleeremans, and P. Wilken, eds. 2009. *Oxford Companion to Consciousness*. New York: Oxford University Press.

Beeckmans, J. 2007. Can higher-order representation theories pass scientific muster? *Journal of Consciousness Studies* 14 (9–10):90–111.

Behne, T., M. Carpenter, J. Call, and M. Tomasello. 2005. Unwilling versus unable: Infants' understanding of intentional action. *Developmental Psychology* 41:328–337.

Bekoff, M., and C. Allen. 1997. Cognitive ethology: Slayers, skeptics, and proponents. In *Anthropomorphism, Anecdotes, and Animals*, ed. R. Mitchell, N. Thompson, and H. Miles. Albany: SUNY Press.

Bekoff, M., C. Allen, and G. Burghardt, eds. 2002. *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition*. Cambridge, MA: MIT Press.

Bennett, J. 1964. *Rationality*. London: Routledge & Kegan Paul.

Bennett, J. 1966. *Kant's Analytic*. Cambridge: Cambridge University Press.

Bennett, J. 1988. Thoughtful brutes. *Proceedings and Addresses of the American Philosophical Association* 62:197–210.

Bennett, J. 1991. Folk-psychological explanations. In *The Future of Folk Psychology*, ed. J. Greenwood. Cambridge: Cambridge University Press.

Bermúdez, J. 1995. Nonconceptual content: From personal experience to subpersonal computational states. *Mind and Language* 10:333–369.

Bermúdez, J. 1998. *The Paradox of Self-Consciousness*. Cambridge, MA: MIT Press.

Bermúdez, J. 2003. *Thinking without Words*. New York: Oxford University Press.

Bermúdez, J. 2007a. What is at stake in the debate about nonconceptual content? In *Philosophical Perspectives: Philosophy of Mind*, vol. 21, ed. J. Hawthorne. Malden, MA: Blackwell.

Bermúdez, J. 2007b. The object properties model of object perception: Between the binding model and the theoretical model. In *The Interplay between Consciousness and Concepts*, ed. R. Gennaro. Exeter: Imprint Academic.

Bermúdez, J. 2009. Mindreading in the animal kingdom. In *The Philosophy of Animal Minds*, ed. R. Lurz. Cambridge: Cambridge University Press.

Bermúdez, José, and Arnon Cahen. 2010. Nonconceptual mental content. In *The Stanford Encyclopedia of Philosophy* (spring 2010 edition), ed. Edward N. Zalta, http://plato.stanford.edu/archives/spr2010/entries/content-nonconceptual.

Beshkar, M. 2008. Animal consciousness. *Journal of Consciousness Studies* 15 (3):5–33.

Birch, S., and P. Bloom. 2004. Understanding children's and adults' limitations in mental state reasoning. *Trends in Cognitive Sciences* 8:255–260.

Blackmore, S. 2004. *Consciousness: An Introduction*. Oxford: Oxford University Press.

Block, N. 1986. Advertisement for a semantics for psychology. In *Midwest Studies in Philosophy*, vol. 10, *Studies in the Philosophy of Mind*, ed. P. French, T. Uehling, and H. Wettstein. Minneapolis: University of Minnesota Press.

Block, N. 1990. Inverted earth. In *Philosophical Perspectives, 4*, ed. J. Tomberlin. Atascadero, CA: Ridgeview.

Block, N. 1995. On a confusion about the function of consciousness. *Behavioral and Brain Sciences* 18:227–247.

Block, N. 1996. Mental paint and mental latex. In *Perception*, ed. E. Villanueva. Atascadero, CA: Ridgeview.

Block, N. 2007. Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences* 30:481–499.

Block, N. 2011. The higher-order approach to consciousness is defunct. *Analysis* 71.

Block, N., O. Flanagan, and G. Güzeldere, eds. 1997. *The Nature of Consciousness*. Cambridge, MA: MIT Press.

Block, N., and R. Stalnaker. 1999. Conceptual analysis, dualism, and the explanatory gap. *Philosophical Review* 108:1–46.

Bloom, P. 2000. *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.

Bloom, P. 2001. Precis of "How Children Learn the Meanings of Words." *Behavioral and Brain Sciences* 24:1095–1134.

Bloom, P., and T. German. 2000. Two reasons to abandon the false belief task as a test of theory of mind. *Cognition* 77:B25–B31.

Bonnar, L., F. Gosselin, and P. Schyns. 2002. Understanding Dali's *Slave Market with the Disappearing Bust of Voltaire*: A case study in the scale information driving perception. *Perception* 31:683–691.

Botterell, A. 2001. Conceiving what is not there. *Journal of Consciousness Studies* 8 (8):21–42.

Boucher, J. 2001. "Lost in a sea of time": Time-parsing and autism. In *Time and Memory*, ed. C. Hoerl and T. McCormack. New York: Oxford University Press.

Bozoki, A., M. Grossman, and E. Smith. 2006. Can patients with Alzheimer's disease learn a category implicitly? *Neuropsychologia* 44:816–827.

Brentano, F. 1874/1973. *Psychology from an Empirical Standpoint*. New York: Humanities.

Brewer, B. 1999. *Perception and Reason*. Oxford: Oxford University Press.

Brewer, B. 2005. Do sense experiential states have conceptual content? In *Contemporary Debates in Epistemology*, ed. E. Sosa and M. Steup. Oxford: Blackwell.

Briscoe, R. 2008. Another look at the two visual systems hypothesis: The argument from illusion studies. *Journal of Consciousness Studies* 15 (8):35–62.

Brook, A. 1994. *Kant and the Mind*. New York: Cambridge University Press.

Brook, A. 2005. Kant, cognitive science, and contemporary neo-Kantianism. *Journal of Consciousness Studies* 11 (10–11):1–25.

Brook, A., and P. Raymont. 2006. The representational base of consciousness. *Psyche* 12 (2). http://theassc.org/files/assc/2638.pdf.

Brook, A., and P. Raymont. 2010. The unity of consciousness. In *The Stanford Encyclopedia of Philosophy* (fall 2010 edition), ed. Edward N. Zalta, http://plato.stanford.edu/archives/fall2010/entries/consciousness-unity.

Brown, C. 2008. Narrow mental content. In *The Stanford Encyclopedia of Philosophy* (winter 2008 edition), ed. Edward N. Zalta, http://plato.stanford.edu/archives/win2008/entries/content-narrow.

Browne, D. 2004. Do dolphins know their own minds? *Biology and Philosophy* 19:633–653.

Bullier, J. 2001. Feedback connections and conscious vision. *Trends in Cognitive Sciences* 9:369–370.

Buras, T. 2009. An argument against causal theories of mental content. *American Philosophical Quarterly* 46:117–129.

Buttelmann, D., M. Carpenter, and M. Tomasello. 2009. Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition* 112:337–342.

Butterfill, S. 2009. Seeing causes and hearing gestures. *Philosophical Quarterly* 59: 405–428.

Byrne, A. 1997. Some like it HOT: Consciousness and higher-order thoughts. *Philosophical Studies* 86:103–129.

Byrne, A. 2001. Intentionalism defended. *Philosophical Review* 110:199–240.

Byrne, A. 2004. What phenomenal consciousness is like. In *Higher-Order Theories of Consciousness: An Anthology*, ed. R. Gennaro. Amsterdam: John Benjamins.

Byrne, A. 2005. Perception and conceptual content. In *Contemporary Debates in Epistemology*, ed. E. Sosa and M. Steup. Oxford: Blackwell.

Byrne, A. 2010. Inverted qualia. In *The Stanford Encyclopedia of Philosophy* (spring 2010 edition), ed. Edward N. Zalta, http://plato.stanford.edu/archives/spr2010/entries/qualia-inverted.

Byrne, A., and M. Tye. 2006. Qualia ain't in the head. *Noûs* 40:241–255.

Campbell, J. 2002. *Reference and Consciousness*. New York: Oxford University Press.

Campbell, J. 2006a. Does visual reference depend on sortal classification? Reply to Clark. *Philosophical Studies* 127:221–237.

Campbell, J. 2006b. What is the role of location in the sense of a visual demonstrative? Reply to Matthen. *Philosophical Studies* 127:239–254.

Carey, S. 1985. *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

Carey, S. 2000. The origin of concepts. *Journal of Cognition and Development* 1:37–41.

Carey, S. 2009. *The Origin of Concepts*. New York: Oxford University Press.

Carey, S., and E. Spelke. 1996. Science and core knowledge. *Philosophy of Science* 63:515–533.

Carruthers, P. 1989. Brute experience. *Journal of Philosophy* 86:258–269.

Carruthers, P. 1996. Simulation and self-knowledge: a defence of theory-theory. In *Theories of Theories of Mind*, ed. P. Carruthers and P. Smith. New York: Cambridge University Press.

Carruthers, P. 1998. Natural theories of consciousness. *European Journal of Philosophy* 6:203–222.

Carruthers, P. 1999. Sympathy and subjectivity. *Australasian Journal of Philosophy* 77:465–482.

Carruthers, P. 2000. *Phenomenal Consciousness*. Cambridge: Cambridge University Press.

Carruthers, P. 2004. HOP over FOR, HOT theory. In *Higher-Order Theories of Consciousness: An Anthology*, ed. R. Gennaro. Amsterdam: John Benjamins.

Carruthers, P. 2005. *Consciousness: Essays from a Higher-Order Perspective*. New York: Oxford University Press.

Carruthers, P. 2008. Meta-cognition in animals: A skeptical look. *Mind and Language* 23:58–89.

Carruthers, P. 2009a. Invertebrate concepts confront the generality constraint (and win). In *The Philosophy of Animal Minds*, ed. R. Lurz. Cambridge: Cambridge University Press.

Carruthers, P. 2009b. How we know our own minds: The relationship between mind-reading and metacognition. *Behavioral and Brain Sciences* 32:121–138.

Carruthers, P., and P. Smith, eds. 1996. *Theories of Theories of Mind*. New York: Cambridge University Press.

Carruthers, P., S. Laurence, and S. Stich, eds. 2005. *The Innate Mind: Structure and Contents*. New York: Oxford University Press.

Carruthers, P., S. Laurence, and S. Stich, eds. 2007. *The Innate Mind: Foundations and the Future*. New York: Oxford University Press.

Carruthers, P., and B. Veillet. 2007. The phenomenal concept strategy. In *The Interplay between Consciousness and Concepts*, ed. R. Gennaro. Exeter: Imprint Academic.

Carver, L., and P. Bauer. 1999. When the event is more than the sum of its parts: Long-term recall of event sequences by 9-month-old infants. *Memory* 7:147–174.

Carver, L., and P. Bauer. 2001. The dawning of a past: The emergence of long-term explicit memory in infancy. *Journal of Experimental Psychology: General* 130:726–745.

Caston, V. 2002. Aristotle on consciousness. *Mind* 111:751–815.

Cavanna, A., and A. Nani. 2008. Do consciousness and attention have shared neural correlates? *Psyche* 14, http://theassc.org/files/assc/2262.pdf.

Chalmers, D. 1993. Connectionism and compositionality: Why Fodor and Pylyshyn were wrong. *Philosophical Psychology* 6:305–319.

Chalmers, D. 1995. Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2:200–219.

Chalmers, D. 1996. *The Conscious Mind*. Oxford: Oxford University Press.

Chalmers, D. 1999. Materialism and the metaphysics of modality. *Philosophy and Phenomenological Research* 59:473–497.

Chalmers, D. 2000. What is a neural correlate of consciousness? In *Neural Correlates of Consciousness: Empirical and Conceptual Questions*, ed. T. Metzinger. Cambridge, MA: MIT Press.

Chalmers, D., ed. 2002. *Philosophy of Mind: Classical and Contemporary Readings*. New York: Oxford University Press.

Chalmers, D. 2003. The nature of narrow content. *Philosophical Issues* 13:46–66.

Chalmers, D. 2004. The representational character of experience. In *The Future for Philosophy*, ed. B. Leiter. Oxford: Oxford University Press.

Chalmers, D. 2007. Phenomenal concepts and the explanatory gap. In *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, ed. T. Alter and S. Walter. New York: Oxford University Press.

Chalmers, D. 2010. *The Character of Consciousness*. New York: Oxford University Press.

Chalmers, D., and F. Jackson. 2001. Conceptual analysis and reductive explanation. *Philosophical Review* 110:315–361.

Chen, X. 2007. The object bias and the study of scientific revolutions: Lessons from developmental psychology. *Philosophical Psychology* 20:479–503.

Chuard, P. 2006. Demonstrative concepts without re-identification. *Philosophical Studies* 130:153–201.

Chuard, P. 2007. The riches of experience. In *The Interplay between Consciousness and Concepts*, ed. R. Gennaro. Exeter: Imprint Academic.

Churchland, P. S. 1986. *Neurophilosophy*. Cambridge, MA: MIT Press.

Churchland, P. S. 2002. *Brain-Wise*. Cambridge, MA: MIT Press.

Clark, Andy. 2001. Visual experience and motor action: Are the bonds too tight? *Philosophical Review* 110:495–519.

Clark, Austen. 2000. *A Theory of Sentience*. Oxford: Oxford University Press.

Clark, Austen. 2006. Attention and inscrutability: A commentary on John Campbell, *Reference and Consciousness*. *Philosophical Studies* 127:167–193.

Clayton, N., T. Bussey, and A. Dickinson. 2003. Can animals recall the past and plan for the future? *Nature Reviews: Neuroscience* 4:685–691.

Clayton, N., N. Emery, and A. Dickinson. 2006. The rationality of animal memory: Complex caching strategies of western scrub jays. In *Rational Animals?* ed. S. Hurley and M. Nudds. New York: Oxford University Press.

Cleeremans, A., ed. 2003. *The Unity of Consciousness: Binding, Integration, and Dissociation*. Oxford: Oxford University Press.

Cleeremans, A., A. Destrebecqz, and M. Boyer. 1998. Implicit learning: News from the front. *Trends in Cognitive Sciences* 2:406–416.

Cleeremans, A., and L. Jiménez. 2002. Implicit learning and consciousness: A graded, dynamic perspective. In *Implicit Learning and Consciousness*, ed. R. French and A. Cleeremans. Psychology Press.

Cleeremans, A., B. Timmermans, and A. Pasquali. 2007. Consciousness and metarepresentation: A computational sketch. *Neural Networks* 20:1032–1039.

Coliva, A. 2003. The argument from the finer-grained content of colour experiences: A redefinition of its role within the debate between McDowell and non-conceptual theorists. *Dialectica* 57:57–70.

Cowey, A. and V. Walsh. 2000. Magnetically induced phosphenes in sighted, blind and blindsighted subjects. *NeuroReport* 11:3269–3273.

Cowie, F. 1999. *What's Within? Nativism Reconsidered*. New York: Oxford University Press.

Crick, F. 1994. *The Astonishing Hypothesis: The Scientific Search for the Soul*. New York: Scribners.

Crick, F., and C. Koch. 1990. Toward a neurobiological theory of consciousness. *Seminars in Neuroscience* 2:263–275.

Crick, F., and C. Koch. 1995. Cortical areas in visual awareness. *Nature* 377:294–295.

Crowther, T. 2006. Two conceptions of conceptualism and nonconceptualism. *Erkenntnis* 65:245–276.

Cundall, M. 2008. Autism. In *Internet Encyclopedia of Philosophy*, http://www.iep.utm.edu/autism.

Dainton, B. 2000. *Stream of Consciousness*. New York: Routledge.

Dainton, B. 2007. Coming together: The unity of conscious experience. In *The Blackwell Companion to Consciousness*, ed. M. Velmans and S. Schneider. Malden, MA: Blackwell.

Damasio, A. 1999. *The Feeling of What Happens*. Harcourt.

Damasio, H., D. Tranel, T. Grabowski, R. Adolphs, and A. Damasio. 2004. Neural systems behind word and concept retrieval. *Cognition* 92:179–229.

Day, R., and B. McKenzie. 1981. Infant perception of the invariant size of approaching and receding objects. *Developmental Psychology* 37:576–586.

de Gardelle, V., J. Sackur, and S. Kouider. 2009. Perceptual illusions in brief visual presentations. *Consciousness and Cognition* 18:569–577.

DeGrazia, D. 2009. Self-awareness in animals. In *The Philosophy of Animal Minds*, ed. R. Lurz. Cambridge: Cambridge University Press.

Dehaene, S., J. Changeux, L. Nacchache, J. Sackut, and C. Sergent. 2006. Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences* 10:204–211.

Del Cul, A., S. Baillet, and S. Dehaene. 2007. Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS Biology* 5:2408–2423.

Dennett, D. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.

Dennett, D. 1988. Quining qualia. In *Consciousness and Contemporary Science*, ed. A. Marcel and E. Bisiach. New York: Oxford University Press.

Dennett, D. 1991. *Consciousness Explained*. Boston: Little, Brown.

Dennett, D. 2005. *Sweet Dreams*. Cambridge, MA: MIT Press.

De Quincey, C. 2006. Switched-on consciousness. *Journal of Consciousness Studies* 13 (4):7–12.

Dere, E., E. Kart-Teke, J. Huston, and D. Silva. 2006. The case for episodic memory in animals. *Neuroscience and Biobehavioral Reviews* 30:1206–1224.

Diaz-Leon, E. 2008. Defending the phenomenal concept strategy. *Australasian Journal of Philosophy* 86:597–610.

Dokic, J., and E. Pacherie. 2001. Shades and concepts. *Analysis* 61:193–202.

Dretske, F. 1981. *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Dretske, F. 1988. Sensation and perception. In *Perceptual Knowledge*, ed. J. Dancy. Oxford: Oxford University Press.

Dretske, F. 1993. Conscious experience. *Mind* 102:263–283.

Dretske, F. 1995. *Naturalizing the Mind*. Cambridge, MA: MIT Press.

Dretske, F. 2004. Change blindness. *Philosophical Studies* 120:1–18.

Dretske, F. 2007. What change blindness teaches about consciousness. In *Philosophical Perspectives: Philosophy of Mind*, vol. 21, ed. J. Hawthorne. Malden, MA: Blackwell.

Droege, P. 2003. *Caging the Beast*. Philadelphia: John Benjamins.

Earl, D. 2007. Concepts. In *Internet Encyclopedia of Philosophy*, http://www.iep.utm.edu/concepts.

Edelman, G. 1989. *The Remembered Present: A Biological Theory of Consciousness*. New York: Basic Books.

Edelman, G., B. Baars, and A. Seth. 2005. Identifying hallmarks of consciousness in non-mammalian species. *Consciousness and Cognition* 14:169–187.

Edelman, G., and A. Seth. 2009. Animal consciousness: A synthetic approach. *Trends in Neurosciences* 32:476–484.

Edelman, G., and G. Tononi. 2000a. Reentry and the dynamic core: Neural correlates of conscious experience. In *Neural Correlates of Consciousness: Empirical and Conceptual Questions*, ed. T. Metzinger. Cambridge, MA: MIT Press.

Edelman, G., and G. Tononi. 2000b. *A Universe of Consciousness*. New York: Basic Books.

Eichenbaum, H., N. Fortin, C. Ergorul, S. Wright, and K. Agster. 2005. Episodic recollection in animals: "If it walks like a duck and quacks like a duck . . ." *Learning and Motivation* 36:190–207.

Eilan, N., C. Hoerl, T. McCormack, and J. Roessler, eds. 2005. *Joint Attention: Communication and Other Minds*. New York: Oxford University Press.

Emery, N., and N. Clayton. 2001. Effects of experience and social context on prospective caching strategies in scrub jays. *Nature* 414:443–446.

Emery, N., and N. Clayton. 2009. Comparative social cognition. *Annual Review of Psychology* 60:87–113.

Evans, G. 1982. *The Varieties of Reference*. Oxford: Oxford University Press.

Fahrenfort, J., H. Scholte, and V. Lamme. 2007. Masking disrupts reentrant processing in human visual cortex. *Journal of Cognitive Neuroscience* 19:1488–1497.

Farah, M. 2004. *Visual Agnosia*. 2nd ed. Cambridge, MA: MIT Press.

Farrant, A., J. Boucher, and M. Blades. 1999. Metamemory in children with autism. *Child Development* 70:107–131.

Feinberg, T. 2000. The nested hierarchy of consciousness: A neurobiological solution to the problem of mental unity. *Neurocase* 6:75–81.

Feinberg, T. 2001. *Altered Egos: How the Brain Creates the Self*. New York: Oxford University Press.

Feinberg, T. 2009. *From Axons to Identity: Neurological Explorations of the Nature of the Self*. New York: W. W. Norton.

Fivaz-Depeursinge, E., N. Favez, and F. Frascarolo. 2004. Threesome intersubjectivity in infancy: A contribution to the development of self-awareness. In *The Structure and Development of Self-Consciousness*, ed. D. Zahavi, T. Grünbaum, and J. Parnas. Amsterdam: John Benjamins.

Fitzpatrick, S. 2008. Doing away with Morgan's canon. *Mind and Language* 23: 224–246.

Fitzpatrick, S. 2009. The primate mindreading controversy: A case study in simplicity and methodology in animal psychology. In *The Philosophy of Animal Minds*, ed. R. Lurz. Cambridge: Cambridge University Press.

Flanagan, O. 1992. *Consciousness Reconsidered*. Cambridge, MA: MIT Press.

Flohr, H. 1995. An information processing theory of anaesthesia. *Neuropsychologia* 33:1169–1180.

Flohr, H. 2000. NMDA receptor-mediated computational processes and phenomenal consciousness. In *Neural Correlates of Consciousness: Empirical and Conceptual Questions*, ed. T. Metzinger. Cambridge, MA: MIT Press.

Flombaum, J., and L. Santos. 2005. Rhesus monkeys attribute perceptions to others. *Current Biology* 15:447–452.

Fodor, J. 1974. Special sciences. *Synthese* 28:77–115.

Fodor, J. 1975. *The Language of Thought*. Cambridge, MA: Harvard University Press.

Fodor, J. 1981. *Representations*. Cambridge, MA: MIT Press.

Fodor, J. 1983. *Modularity of Mind*. Cambridge, MA: MIT Press.

Fodor, J. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Fodor, J. 1990. A theory of content. In J. Fodor, *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Fodor, J. 1991. A modal argument for narrow content. *Journal of Philosophy* 88:5–25.

Fodor, J. 1998. *Concepts: Where Cognitive Science Went Wrong*. New York: Oxford University Press.

Fodor, J. 2007. The revenge of the given. In *Contemporary Debates in the Philosophy of Mind*, ed. B. McLaughlin and J. Cohen. Oxford: Blackwell.

Fodor, J., and Z. Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition* 28:3–71.

Ford, J., and D. W. Smith. 2006. Consciousness, self, and attention. In *Self-Representational Approaches to Consciousness*, ed. U. Kriegel and K. Williford. Cambridge, MA: MIT Press.

Frensch, P., and D. Runger. 2003. Implicit learning. *Current Directions in Psychological Science* 12:13–17.

Frith, C. 1992. *The Cognitive Neuropsychology of Schizophrenia*. East Sussex: Psychology Press.

Frith, C., and F. Happé. 1999. Theory of mind and self-consciousness: What is it like to be autistic? *Mind and Language* 14:1–22.

Frith, C., and E. Hill, eds. 2003. *Autism: Mind and Brain*. New York: Oxford University Press.

Gaillard, R., S. Dehaene, C. Adam, S. Clemenceau, D. Hasboun, M. Baulac, L. Cohen, and L. Naccache. Converging intracranial markers of conscious access. *Public Library of Science Biology* 7:472–492.

Gallace, A., and C. Spence. 2010. Touch and the body: The role of the somatosensory cortex in tactile awareness. *Psyche* 16:31–67.

Gallagher, S. 2004. Agency, ownership, and alien control in schizophrenia. In *The Structure and Development of Self-Consciousness*, ed. D. Zahavi, T. Grünbaum, and J. Parnas. Amsterdam: John Benjamins.

Gallagher, S. 2005. *How the Body Shapes the Mind*. Oxford: Oxford University Press.

Gallup, G. 1970. Chimpanzees: Self-recognition. *Science* 167:86–87.

Garson, James. 2010. Connectionism. In *The Stanford Encyclopedia of Philosophy* (fall 2010 edition), ed. Edward N. Zalta, http://plato.stanford.edu/archives/fall2010/entries/connectionism.

Gelman, S. 2003. *The Essential Child*. New York: Oxford University Press.

Gendler, T., and J. Hawthorne, eds. 2002. *Conceivability and Possibility*. New York: Oxford University Press.

Gendler, T., and J. Hawthorne, eds. 2006. *Perceptual Experience*. New York: Oxford University Press.

Gennaro, R. 1992. Consciousness, self-consciousness, and episodic memory. *Philosophical Psychology* 5:333–347.

Gennaro, R. 1993. Brute experience and the higher-order thought theory of consciousness. *Philosophical Papers* 22:51–69.

Gennaro, R. 1995. Does mentality entail consciousness? *Philosophia* 24:331–358.

Gennaro, R. 1996. *Consciousness and Self-Consciousness: A Defense of the Higher-Order Thought Theory of Consciousness*. Amsterdam: John Benjamins.

Gennaro, R. 1999. Leibniz on consciousness and self-consciousness. In *New Essays on the Rationalists*, ed. R. Gennaro and C. Huenemann. New York: Oxford University Press.

Gennaro, R. 2002. Jean-Paul Sartre and the HOT theory of consciousness. *Canadian Journal of Philosophy* 32:293–330.

Gennaro, R. 2003. Papineau on the actualist HOT theory of consciousness. *Australasian Journal of Philosophy* 81:581–586.

Gennaro, R. 2004a. Higher-order thoughts, animal consciousness, and misrepresentation: A reply to Carruthers and Levine. In *Higher-Order Theories of Consciousness: An Anthology*, ed. R. Gennaro. Amsterdam: John Benjamins.

Gennaro, R., ed. 2004b. *Higher-Order Theories of Consciousness: An Anthology*. Amsterdam: John Benjamins.

Gennaro, R. 2005a. Consciousness. In *Internet Encyclopedia of Philosophy*, http://www.iep.utm.edu/consciou.

Gennaro, R. 2005b. The HOT theory of consciousness: Between a rock and a hard place? *Journal of Consciousness Studies* 12 (2):3–21.

Gennaro, R. 2006a. Between pure self-referentialism and the (extrinsic) HOT theory of consciousness. In *Self-Representational Approaches to Consciousness*, ed. U. Kriegel and K. Williford. Cambridge, MA: MIT Press.

Gennaro, R. 2006b. Review of Peter Carruthers' *Consciousness: Essays from a Higher-Order Perspective*. *Psyche* 12, http://theassc.org/files/assc/2645.pdf.

Gennaro, R., ed. 2007. *The Interplay between Consciousness and Concepts*. Exeter, UK: Imprint Academic. Also a special double issue of the *Journal of Consciousness Studies* 14 (9–10).

Gennaro, R. 2008a. Representationalism, peripheral awareness, and the transparency of experience. *Philosophical Studies* 139:39–56.

Gennaro, R. 2008b. Are there pure conscious events? In *Revisiting Mysticism*, ed. C. Chakrabarti and G. Haist. Newcastle: Cambridge Scholars Press.

Gennaro, R. 2009. Animals, consciousness, and I-thoughts. In *Philosophy of Animal Minds*, ed. Robert Lurz. New York: Cambridge University Press.

Gennaro, R., D. Herrmann, and M. Sarapata. 2006. Aspects of the unity of consciousness and everyday memory failures. *Consciousness and Cognition* 15:372–385.

Georgalis, N. 2006. *The Primacy of the Subjective*. Cambridge, MA: MIT Press.

Gerken, M. 2008. Is there a simple argument for higher-order representation theories of awareness consciousness? *Erkenntnis* 69:243–259.

Ginsborg, H. 2006. Empirical concepts and the content of experience. *European Journal of Philosophy* 14:349–372.

Ginsborg, H. 2008. Was Kant a nonconceptualist? *Philosophical Studies* 137:65–77.

Gois, I. 2010. A dilemma for higher-order theories of consciousness. *Philosophia* 38:143–156.

Goldberg, I., M. Harel, and R. Malach. 2006. When the brain loses its self: Prefrontal inactivation during sensorimotor processing. *Neuron* 50:329–339.

Goldman, A. 1993. Consciousness, folk psychology, and cognitive science. *Consciousness and Cognition* 2:264–282.

Goldman, A. 2006. *Simulating Minds*. New York: Oxford University Press.

Goldstone, R., and L. Barsalou. 1998. Reuniting perception and conception. *Cognition* 65:231–262.

Goldstone, R. and Hendrickson, A. 2009. Categorical perception. *WIREs Cognitive Science*, http://onlinelibrary.wiley.com/doi/10.1002/wcs.26/pdf.

Gómez, J. 2005. Joint attention and the notion of subject: Insights from apes, normal children, and children with autism. In *Joint Attention: Communication and Other Minds*, ed. N. Eilan, C. Hoerl, T. McCormack, and J. Roessler. New York: Oxford University Press.

Goodale, M. 2007. Duplex vision: Separate cortical pathways for conscious perception and the control of action. In *The Blackwell Companion to Consciousness*, ed. M. Velmans and S. Schneider. Malden, MA: Blackwell.

Gopnik, A. 1996. The scientist as child. *Philosophy of Science* 63:485–514.

Gopnik, A. 2009. *The Philosophical Baby*. New York: Farrar, Straus & Giroux.

Gopnik, A., and A. Meltzoff. 1997. *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.

Graham, G., T. Horgan, and J. Tienson. 2007. Consciousness and intentionality. In *The Blackwell Companion to Consciousness*, ed. M. Velmans and S. Schneider. Malden, MA: Blackwell

Grandin, T. 1995. *Thinking in Pictures*. New York: Vintage Press.

Grandy, Richard E. 2008. Sortals. In *The Stanford Encyclopedia of Philosophy* (fall 2008 edition), ed. Edward N. Zalta, http://plato.stanford.edu/archives/fall2008/entries/sortals.

Griffin, D., and G. Speck. 2004. New evidence of animal consciousness. *Animal Cognition* 7:5–18.

Griffiths, P. 2002. What is innateness? *Monist* 85:70–85.

Griffiths, P., and E. Machery. 2008. Innateness, canalization, and "biologicizing the mind." *Philosophical Psychology* 21:397–414.

Grill-Spector, K., and R. Malach. 2004. The human visual cortex. *Annual Review of Neuroscience* 7:649–677.

Grossenbacher, P., ed. 2001. *Finding Consciousness in the Brain*. Amsterdam: John Benjamins.

Gunther, Y., ed. 2003. *Essays on Nonconceptual Content*. Cambridge, MA: MIT Press.

Güzeldere, G. 1995. Is consciousness the perception of what passes in one's own mind? In *Conscious Experience*, ed. T. Metzinger. Paderborn: Ferdinand Schöningh.

Hampton, R. 2005. Can rhesus monkeys discriminate between remembering and forgetting? In *The Missing Link in Cognition: Origins of Self-Reflective Consciousness*, ed. H. Terrace and J. Metcalfe. New York: Oxford University Press.

Hanna, R. 2005. Kant and nonconceptual content. *European Journal of Philosophy* 13:247–290.

Hanna, R. 2008. Kantian non-conceptualism. *Philosophical Studies* 137:41–64.

Hardcastle, V. 1995. *Locating Consciousness*. Amsterdam: John Benjamins.

Hardcastle, V. 2004. HOT theories of consciousness: More sad tales of philosophical intuitions gone astray. In *Higher-Order Theories of Consciousness: An Anthology*, ed. R. Gennaro. Amsterdam: John Benjamins.

Hare, B., J. Call, B. Agnetta, and M. Tomasello. 2000. Chimpanzees know what conspecifics do and do not see. *Animal Behaviour* 59:771–785.

Hare, B., J. Call, and M. Tomasello. 2001. Do chimpanzees know what conspecifics know? *Animal Behaviour* 61:139–151.

Hare, B., and M. Tomasello. 2004. Chimpanzees are more skilled in competitive than in cooperative cognitive tasks. *Animal Behaviour* 68:571–581.

Harman, G. 1973. *Thought*. Princeton, NJ: Princeton University Press.

Harman, G. 1990. The intrinsic quality of experience. In *Philosophical Perspectives*, vol. 4, ed. J. Tomberlin. Atascadero, CA: Ridgeview.

Hassin, R., J. Bargh, A. Engell, and K. McCulloch. 2009. Implicit working memory. *Consciousness and Cognition* 18:665–678.

Hasson, U., Y. Nir, I. Levy, G. Fuhrmann, and R. Malach. 2004. Intersubject synchronization of cortical activity during natural vision. *Science* 303:1634–1640.

Hauser, M., P. MacNeilage, and M. Ware. 1996. Numerical representations in primates. *Proceedings of the National Academy of Science USA* 93:1514–1517.

Hauser, M., and L. Santos. 2007. The evolutionary ancestry of our knowledge of tools: from percepts to concepts. In *Creations of the Mind*, ed. E. Margolis and S. Laurence. New York: Oxford University Press.

Hawthorne, J. 1989. On the compatibility of connectionist and classical models. *Philosophical Psychology* 2:5–15.

Hayne, H. 2007. Infant memory development: new questions, new answers. In *Short- and Long-Term Memory in Infancy and Early Childhood*, ed. L. Oakes and P. Bauer. New York: Oxford University Press.

Heck, R. 2000. Non-conceptual content and the "space of reasons." *Philosophical Review* 109:483–523.

Heck, R. 2007. Are there different kinds of content? In *Contemporary Debates in the Philosophy of Mind*, ed. B. McLaughlin and J. Cohen. Oxford: Blackwell.

Hellie, B. 2007. Higher-order intentionality and higher-order acquaintance. *Philosophical Studies* 134:289–324.

Hill, C. 1991. *Sensations*. Cambridge, MA: Cambridge University Press.

Hill, C. 1997. Imaginability, conceivability, possibility, and the mind–body problem. *Philosophical Studies* 87:61–85.

Hill, C., and B. McLaughlin. 1998. There are fewer things in reality than are dreamt of in Chalmers' philosophy. *Philosophy and Phenomenological Research* 59:445–454.

Hillier, A., and L. Allinson. 2002. Understanding embarrassment among those with autism: Breaking down the complex emotion of embarrassment among those with autism. *Journal of Autism and Developmental Disorders* 32:583–592.

Hochstein, S., and M. Ahissar. 2002. View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron* 36:791–804.

Hoerl, C. 2009. Time and tense in perceptual experience. *Philosopher's Imprint* 9:1–18.

Hohwy, J. 2007. The search for neural correlates of consciousness. *Philosophy Compass* 2–3:461–474.

Hohwy, J. 2009. The neural correlates of consciousness: New experimental approaches needed? *Consciousness and Cognition* 18:428–438.

Horgan, T. 1984. Jackson on physical information and qualia. *Philosophical Quarterly* 34:147–152.

Horgan, T., and J. Tienson. 2002. The intentionality of phenomenology and the phenomenology of intentionality. In *Philosophy of Mind: Classical and Contemporary Readings*, ed. D. Chalmers. New York: Oxford University Press.

Hossack, K. 2002. Self-knowledge and consciousness. *Proceedings of the Aristotelian Society* 102:163–181.

Hossack, K. 2003. Consciousness in act and action. *Phenomenology and the Cognitive Sciences* 2:187–203.

Humphreys, G. 2003. Conscious visual representations built from multiple binding processes: Evidence from neuropsychology. In *The Unity of Consciousness: Binding, Integration, and Dissociation*, ed. A. Cleeremans. Oxford: Oxford University Press.

Hurlburt, R., F. Happé, and U. Frith. 1994. Sampling the form of inner experience in three adults with Asperger Syndrome. *Psychological Medicine* 24:385–395.

Hurley, S. 2001. Overintellectualizing the mind. *Philosophy and Phenomenological Research* 63:423–431.

Hurley, S. 2003. Action, the unity of consciousness, and vehicle externalism. In *The Unity of Consciousness: Binding, Integration, and Dissociation*, ed. A. Cleeremans. Oxford: Oxford University Press.

Hurley, S., and M. Nudds, eds. 2006. *Rational Animals?* New York: Oxford University Press.

Husserl, E. 1913/1931. *Ideas: General Introduction to Pure Phenomenology [Ideen au einer reinen Phänomenologie und phänomenologischen Philosophie]*. Trans. W. Boyce Gibson. New York: Macmillan.

Jackendoff, R. 1987. *Consciousness and the Computational Mind*. Cambridge, MA: MIT Press.

Jackson, F. 1982. Epiphenomenal qualia. *Philosophical Quarterly* 32:127–136.

Jackson, F. 1986. What Mary didn't know. *Journal of Philosophy* 83:291–295.

Jackson, F. 2004. Postscripts. In *There's Something About Mary*, ed. P. Ludlow, Y. Nagasawa, and D. Stoljar. Cambridge, MA: MIT Press.

Jacoby, L., D. Lindsay, and J. Toth. 1992. Unconscious influences revealed: Attention, awareness, and control. *American Psychologist* 47:802–809.

James, W. 1890. *The Principles of Psychology*. New York: Henry Holt.

Janzen, G. 2005. Self-consciousness and phenomenal character. *Dialogue* 44:707–733.

Janzen, G. 2008. *The Reflexive Nature of Consciousness*. Amsterdam: John Benjamins.

Jeannerod, M. 2004. From self-recognition to self-consciousness. In *The Structure and Development of Self-Consciousness*, ed. D. Zahavi, T. Grünbaum, and J. Parnas. Amsterdam: John Benjamins.

Jehle, D., and U. Kriegel. 2006. An argument against dispositional HOT theory. *Philosophical Psychology* 19:462–476.

Jiang, Y., P. Costello, F. Fang, M. Huang, S. He, and D. Purves. 2006. A gender- and sexual orientation-dependent spatial attentional effect of invisible images. *Proceedings of the National Academy of Sciences of the United States of America* 1003: 17048–17052.

Johnson, S. 2005. Reasoning about intentionality in preverbal infants. In *The Innate Mind: Structure and Contents*, ed. P. Carruthers, S. Laurence, and S. Stich. New York: Oxford University Press.

Kant, I. 1781/1965. *Critique of Pure Reason*. Trans. N. Kemp Smith. New York: Macmillan.

Kastner, S. 2004. Attentional response modulation in the human visual system. In *The Cognitive Neuroscience of Attention*, ed. M. Posner. New York: Guilford Press.

Keenan, J., G. Gallup, and D. Falk. 2003. *The Face in the Mirror*. New York: HarperCollins.

Keil, F. 1989. *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: MIT Press.

Keleman, D., and S. Carey. 2007. The essence of artifacts: Developing the design stance. In *Creations of the Mind*, ed. E. Margolis and S. Laurence. New York: Oxford University Press.

Kellman, P., and E. Spelke. 1983. Perception of partly occluded objects in infancy. *Cognitive Psychology* 15:483–524.

Kelly, S. 2001a. Demonstrative concepts and experience. *Philosophical Review* 110:397–420.

Kelly, S. 2001b. The non-conceptual content of perceptual experience: Situation dependence and fineness of grain. *Philosophy and Phenomenological Research* 62: 601–608.

Kemeny, J., and P. Oppenheim. 1956. On reduction. *Philosophical Studies* 7:6–19.

Kéri, S. 2003. The cognitive neuroscience of category learning. *Brain Research: Brain Research Reviews* 43:85–109.

Khalidi, M. 2007. Innate cognitive capacities. *Mind and Language* 22:92–115.

Khalidi, M. 2009. Should we eliminate the innate? Reply to Griffiths and Machery. *Philosophical Psychology* 22:505–519.

Kidd, C. 2011. Phenomenal consciousness with infallible self-representation. *Philosophical Studies* 152:361–383.

Kihlstrom, J., J. Dorfman, and L. Park. 2007. Implicit and explicit memory and learning. In *The Blackwell Companion to Consciousness*, ed. M. Velmans and S. Schneider. Malden, MA: Blackwell.

Kind, A. 2003. What's so transparent about transparency? *Philosophical Studies* 115:225–244.

Kind, A. 2007. Restrictions on representationalism. *Philosophical Studies* 134:405–427.

Kind, A. 2008. Qualia. In *Internet Encyclopedia of Philosophy*, http://www.iep.utm.edu/qualia.

Kinsbourne, M. 2005. A continuum of self-consciousness that emerges in phylogeny and ontogeny. In *The Missing Link in Cognition: Origins of Self-Reflective Consciousness*, ed. H. Terrace and J. Metcalfe. New York: Oxford University Press.

Kinzler, K., and E. Spelke. 2007. Core systems in human cognition. *Progress in Brain Research* 164:257–264.

Kirk, R. 1994. *Raw Feeling*. New York: Oxford University Press.

Kirk, R. 2005. *Zombies and Consciousness*. New York: Oxford University Press.

Kirk, R. 2009. Zombies. In *The Stanford Encyclopedia of Philosophy* (summer 2009 edition), ed. Edward N. Zalta, http://plato.stanford.edu/archives/sum2009/entries/zombies.

Kitcher, P. 1990. *Kant's Transcendental Psychology*. New York: Oxford University Press.

Kitcher, P. 2010. *Kant's Thinker*. New York: Oxford University Press.

Knowlton, B., and L. Squire. 1993. The learning of categories: Parallel brain systems for item memory and category knowledge. *Science* 262:1747–1749.

Koch, C. 2004. *The Quest for Consciousness: A Neurobiological Approach*. Englewood, CO: Roberts.

Koch, C., and N. Tsuchiya. 2006. Attention and consciousness: Two distinct brain processes. *Trends in Cognitive Sciences* 11:16–22.

Koivisto, M., and A. Revonsuo. 2007. How meaning shapes seeing. *Psychological Science* 18:845–849.

Koriat, A. 2007. Metacognition and consciousness. In *The Cambridge Handbook of Consciousness*, ed. P. Zelazo, M. Moscovitch, and E. Thompson. Cambridge: Cambridge University Press.

Kosslyn, S. 2001. Visual consciousness. In *Finding Consciousness in the Brain*, ed. P. Grossenbacher. Amsterdam: John Benjamins.

Kouider, S., and S. Dehaene. 2007. Levels of processing during non-conscious perception: A critical review of visual masking. *Philosophical Transactions of the Royal Society* 362:857–875.

Kouider, S. 2009. Neurobiological theories of consciousness. In *Encyclopedia of Consciousness*, ed. W. Banks. Oxford: Elsevier.

Kriegel, U. 2002a. PANIC theory and the prospects for a representational theory of phenomenal consciousness. *Philosophical Psychology* 15:55–64.

Kriegel, U. 2002b. Consciousness, permanent self-awareness, and higher-order thought. *Dialogue* 41:517–540.

Kriegel, U. 2003a. Is intentionality dependent upon consciousness? *Philosophical Studies* 116:271–307.

Kriegel, U. 2003b. Consciousness as intransitive self-consciousness: Two views and an argument. *Canadian Journal of Philosophy* 33:103–132.

Kriegel, U. 2003c. Consciousness, higher-order content, and the individuation of vehicles. *Synthese* 134:477–504.

Kriegel, U. 2004a. Consciousness and self-consciousness. *Monist* 87:182–205.

Kriegel, U. 2004b. The functional role of consciousness: A phenomenological approach. *Phenomenology and the Cognitive Sciences* 4:171–193.

Kriegel, U. 2005. Naturalizing subjective character. *Philosophy and Phenomenological Research* 71:23–56.

Kriegel, U. 2006. The same order monitoring theory of consciousness. In *Self-Representational Approaches to Consciousness*, ed. U. Kriegel and K. Williford. Cambridge, MA: MIT Press.

Kriegel, U. 2007a. Philosophical theories of consciousness: Contemporary Western perspectives. In *The Cambridge Handbook of Consciousness*, ed. P. Zelazo, M. Moscovitch, and E. Thompson. Cambridge: Cambridge University Press.

Kriegel, U. 2007b. A cross-order integration hypothesis for the neural correlate of consciousness. *Consciousness and Cognition* 16:897–912.

Kriegel, U. 2008. Real narrow content. *Mind and Language* 23:304–328.

Kriegel, U. 2009a. *Subjective Consciousness*. New York: Oxford University Press.

Kriegel, U. 2009b. Self-representationalism and phenomenology. *Philosophical Studies* 143:357–381.

Kriegel, U. Forthcoming. Personal-level representation. *Protosociology: special issue on "Consciousness and Subjectivity."*

Kriegel, U., and K. Williford, eds. 2006. *Self-Representational Approaches to Consciousness*. Cambridge, MA: MIT Press.

Kripke, S. 1972. *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Krojgaard, P. 2004. A review of object individuation in infancy. *British Journal of Developmental Psychology* 22:159–183.

Lamme, V. 2003. Why visual attention and awareness are different. *Trends in Cognitive Sciences* 7:12–18.

Lamme, V. 2004. Separate neural definitions of visual consciousness and visual attention: A case for phenomenal awareness. *Neural Networks* 17:861–872.

Lamme, V., and P. Roelfsema. 2000. The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences* 23:571–579.

Lau, H., and R. Passingham. 2006. Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences of the United States of America* 103:18763–18768.

Laurence, S., and E. Margolis. 2002. Radical concept nativism. *Cognition* 86:25–55.

Laurier, D. 2004. Nonconceptual contents vs. nonconceptual states. *Grazer Philosophische Studien* 68:23–43.

Leekam, S. 2005. Why do children with autism have a joint attention impairment? In *Joint Attention: Communication and Other Minds*, ed. N. Eilan, C. Hoerl, T. McCormack, and J. Roessler. New York: Oxford University Press.

Legrand, D. 2007a. Pre-reflective self-consciousness: On being bodily in the world. *Janus Head* 9:493–519.

Legrand, D. 2007b. Pre-reflective self-as-subject from experiential and empirical perspectives. *Consciousness and Cognition* 16:583–599.

Le Poidevin, Robin. 2009. The experience and perception of time. In *The Stanford Encyclopedia of Philosophy* (winter 2009 edition), ed. Edward N. Zalta, http://plato.stanford.edu/archives/win2009/entries/time-experience.

Leslie, A. 1984. Infant perception of a manual pick-up event. *British Journal of Developmental Psychology* 2:19–32.

Leslie, A., C. Gallistel, and R. Gelman. 2007. Where integers come from. In *The Innate Mind: Foundations and the Future*, ed. P. Carruthers, S. Laurence, and S. Stich. New York: Oxford University Press.

Levin, D. 2002. Change blindness as visual metacognition. *Journal of Consciousness Studies* 9 (5–6):111–130.

Levine, J. 1983. Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly* 64:354–361.

Levine, J. 1995. Qualia: Intrinsic, relational, or what? In *Conscious Experience*, ed. T. Metzinger. Paderborn: Ferdinand Schöningh.

Levine, J. 2001. *Purple Haze: The Puzzle of Conscious Experience*. Cambridge, MA: MIT Press.

Levine, J. 2006. Awareness and (self-)representation. In *Self-Representational Approaches to Consciousness*, ed. U. Kriegel and K. Williford. Cambridge, MA: MIT Press.

Levine, J. 2007. Anti-materialist arguments and influential replies. In *The Blackwell Companion to Consciousness*, ed. M. Velmans and S. Schnieder. Malden, MA: Blackwell.

Little, M. 1997. Virtue as knowledge: Objections from the philosophy of mind. *Noûs* 31:59–79.

Loar, B. 1990. Phenomenal states. *Philosophical Perspectives* 4:81–108.

Loar, B. 1997. Phenomenal states. [A revised version of Loar 1990.] In *The Nature of Consciousness*, ed. N. Block, O. Flanagan, and G. Güzeldere. Cambridge, MA: MIT Press.

Loar, B. 1999. David Chalmers's *The Conscious Mind*. *Philosophy and Phenomenological Research* 59:465–472.

Loar, B. 2003. Transparent experience and the availability of qualia. In *Consciousness: New Philosophical Perspectives*, ed. Q. Smith and A. Jokic. New York: Oxford University Press.

Locke, J. 1689/1975. *An Essay concerning Human Understanding*. Ed. P. Nidditch. Oxford: Clarendon.

Lormand, E. 1996. Nonphenomenal consciousness. *Noûs* 30:242–261.

Lormand, E. Unpublished. Inner sense until proven guilty.

Lowe, E. 2007. Sortals and the individuation of objects. *Mind and Language* 22:514–533.

Ludlow, P., Y. Nagasawa, and D. Stoljar, eds. 2004. *There's Something about Mary*. Cambridge, MA: MIT Press.

Lurz, R. 2001. Begging the question: A reply to Lycan. *Analysis* 61:313–318.

Lurz, R. 2002. Neither HOT nor COLD: An alternative account of consciousness. *Psyche* 9, http://www.theassc.org/files/assc/2561.pdf.

Lurz, R. 2004. Either FOR or HOR: A false dichotomy. In *Higher-Order Theories of Consciousness: An Anthology*, ed. R. Gennaro. Amsterdam: John Benjamins.

Lurz, R. 2007. In defense of wordless thoughts about thoughts. *Mind and Language* 22:270–296.

Lurz, R. 2009a. Animal minds. In *Internet Encyclopedia of Philosophy*, http://www.iep.utm.edu/ani-mind.

Lurz, R., ed. 2009b. *The Philosophy of Animal Minds*. Cambridge: Cambridge University Press.

Lurz, R. 2009c. If chimpanzees are mindreaders, could behavioral science tell? Toward a solution of the logical problem. *Philosophical Psychology* 22:305–328.

Lurz, R. 2011. *Mindreading Animals*. Cambridge, MA: MIT Press.

Lycan, W. G. 1996. *Consciousness and Experience*. Cambridge, MA: MIT Press.

Lycan, W. G. 2001a. A simple argument for a higher-order representation theory of consciousness. *Analysis* 61:3–4.

Lycan, W. G. 2001b. The case for phenomenal externalism. *Philosophical Perspectives* 15:17–35.

Lycan, W. G. 2004. The superiority of HOP to HOT. In *Higher-Order Theories of Consciousness: An Anthology*, ed. R. Gennaro. Amsterdam: John Benjamins.

Lycan, W. G. 2005. Representational theories of consciousness. In *The Stanford Encyclopedia of Philosophy* (spring 2005 edition), ed. Edward N. Zalta, http://plato.stanford.edu/archives/spr2005/entries/consciousness-representational.

Lycan, W. G. 2008. Phenomenal intentionalities. *American Philosophical Quarterly* 45:233–252.

Lycan, W. G., and Z. Ryder. 2003. The loneliness of the long-distance truck driver. *Analysis* 63:132–136.

Lyyra, P. 2005. Review of José Luis Bermúdez: *Thinking without Words*. *Psyche* 11: http://www.theassc.org/files/assc/2602.pdf.

Lyyra, P. 2009. Two senses for "givenness of consciousness." *Phenomenology and the Cognitive Sciences* 8:67–87.

Machery, E. 2005. Concepts are not a natural kind. *Philosophy of Science* 72:444–467.

Machery, E. 2009. *Doing without Concepts*. New York: Oxford University Press.

Mack, A., and I. Rock. 1998. *Inattentional Blindness*. Cambridge, MA: MIT Press.

MacPherson, F. 2005. Colour inversion problems for representationalism. *Philosophy and Phenomenological Research* 70:127–152.

MacPherson, F. 2006. Ambiguous figures and the content of experience. *Noûs* 40:82–117.

Madole, K., and L. Oakes. 1999. Making sense of infant categorization: Stable proceses and changing representations. *Developmental Review* 19:263–296.

Mandik, P. 2009. Beware of the unicorn. *Journal of Consciousness Studies* 16 (1):5–36.

Mandler, J. 1999. Seeing is not the same as thinking: Commentary on "Making sense of infant categorization." *Developmental Review* 19:297–306.

Mandler, J. 2004. *The Foundations of Mind*. New York: Oxford University Press.

Mandler, J. 2007. The conceptual foundations of animals and artifacts. In *Creations of the Mind*, ed. E. Margolis and S. Laurence. New York: Oxford University Press.

Mandler, J. 2008. On the birth and growth of concepts. *Philosophical Psychology* 21:207–230.

Marcel, A. 1983. Conscious and unconscious perception: Experiments on visual masking and world recognition. *Cognitive Psychology* 15:197–237.

Mareschal, D. 2003. The acquisition and use of implicit categories in early development. In *Early Category and Concept Development*, ed. D. Rakison and L. Oakes. New York: Oxford University Press.

Margolis, E. 1998. How to acquire a concept. *Mind and Language* 13:347–369.

Margolis, E., and S. Laurence, eds. 1999. *Concepts: Core Readings*. Cambridge, MA: MIT Press.

Margolis, E., and S. Laurence. 2007a. The ontology of concepts—abstract objects or mental representations? *Noûs* 41:561–593.

Margolis, E., and S. Laurence, eds. 2007b. *Creations of the Mind*. New York: Oxford University Press.

Margolis, E., and S. Laurence. 2008. Concepts. In *The Stanford Encyclopedia of Philosophy* (fall 2008 edition), ed. Edward N. Zalta, http://plato.stanford.edu/archives/fall2008/entries/concepts.

Marr, D. 1982. *Vision*. New York: W. H. Freeman.

Martin, M. 1992. *Philosophical Review* 101:745–764. [Reprinted in Gunther 2003. Page references are to Gunther 2003.]

Martinez-Conde, S., and S. Macknik. 2008. Magic and the brain. *Scientific American* (December):72–79.

Mashour, G., and E. LaRock. 2008. Inverse zombies, anesthesia awareness, and the hard problem of unconsciousness. *Consciousness and Cognition* 17:1163–1168.

Matey, J. 2006. Two HOTs to handle: The concept of state consciousness in the higher-order thought theory of consciousness. *Philosophical Psychology* 19:151–175.

Matthen, M. 2006. On visual experience of objects: Comments on John Campbell's *Reference and Consciousness*. *Philosophical Studies* 127:195–220.

McDonough, L., and J. Mandler. 1998. Inductive generalization in 9- and 11-month olds. *Developmental Science* 1:227–232.

McDowell, J. 1994. *Mind and World*. Cambridge, MA: Harvard University Press.

McDowell, J. 1998. Having the world in view: Sellars, Kant, and intentionality. *Journal of Philosophy* 95:431–491.

McDowell, J. 2002. Responses. In *Reading McDowell on Mind and World*, ed. N. Smith. New York: Routledge.

McDowell, J. 2008. Responses. In *Experience, Norm, and Nature*, ed. J. Lindgaard. Malden, MA: Blackwell.

McGinn, C. 1989. Can we solve the mind–body problem? *Mind* 98:349–366.

McGinn, C. 1991. *The Problem of Consciousness*. Oxford: Blackwell.

McGinn, C. 1995. Consciousness and space. In *Conscious Experience*, ed. T. Metzinger. Paderborn: Ferdinand Schöningh.

McLaughlin, B., and J. Cohen, eds. 2007. *Contemporary Debates in the Philosophy of Mind*. Oxford: Blackwell.

Medin, D., and A. Ortony. 1989. Psychological essentialism. In *Similarity and Analogical Reasoning*, ed. S. Vosniadou. New York: Cambridge University Press.

Meltzoff, A. 1988. Infant imitation after a 1-week delay: Long-term memory for novel acts and multiple stimuli. *Developmental Psychology* 24:470–476.

Meltzoff, A. 1995. Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology* 31:838–850.

Menzel, C. 2005. Progress in the study of chimpanzee recall and episodic memory. In *The Missing Link in Cognition: Origins of Self-Reflective Consciousness*, ed. H. Terrace and J. Metcalfe. New York: Oxford University Press.

Metcalfe, J., and A. P. Shimamura, eds. 1994. *Metacognition: Knowing about Knowing*. Cambridge, MA: MIT Press.

Metzinger, T., ed. 1995. *Conscious Experience*. Paderborn: Ferdinand Schöningh.

Metzinger, T., ed. 2000. *Neural Correlates of Consciousness: Empirical and Conceptual Questions*. Cambridge, MA: MIT Press.

Michotte, A. 1963. *The Perception of Causality*. London: Methuen.

Miller, E., A. Nieder, D. Freedman, and J. Wallis. 2003. Neural correlates of categories and concepts. *Current Opinion in Neurobiology* 13:198–203.

Millikan, R. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

Milner, A., and M. Goodale. 1995. *The Visual Brain in Action*. Oxford: Oxford University Press.

Mole, C. 2008. Attention and consciousness. *Journal of Consciousness Studies* 15 (4):86–104.

Mole, C. 2009. Attention. In *The Stanford Encyclopedia of Philosophy* (fall 2009 edition), ed. Edward N. Zalta, http://plato.stanford.edu/archives/fall2009/entries/attention.

Montminy, M. 2005. What use is Morgan's canon? *Philosophical Psychology* 18: 399–414.

Moore, G. E. 1903. The refutation of idealism. In *Philosophical Studies*, ed. G. E. Moore. Totowa, NJ: Littlefield, Adams.

Morgan, C. L. 1894. *An Introduction to Comparative Psychology*. London: Walter Scott.

Morin, A. 2006. Levels of consciousness and self-awareness: A comparison and integration of various neurocognitive views. *Consciousness and Cognition* 15: 358–371.

Munkata, Y., and J. McClelland. 2003. Connectionist models of development. *Developmental Science* 6:413–429.

Munkata, Y., J. McClelland, M. Johnson, and R. Siegler. 1997. Rethinking infant knowledge: Toward an adaptive process account of successes and failures in object permanence tasks. *Psychological Review* 104:686–713.

Murphy, G. 2002. *The Big Book of Concepts*. Cambridge, MA: MIT Press.

Nagel, T. 1974. What is it like to be a bat? *Philosophical Review* 83:435–456.

Natsoulas, T. 1993. Consciousness(4): Varieties of intrinsic theory. *Journal of Mind and Behavior* 14:107–132.

Neander, K. 1998. The division of phenomenal labor: A problem for representational theories of consciousness. *Philosophical Perspectives* 12:411–434.

Needham, A. 2001. Object recognition and object segregation in 4.5-month-old infants. *Journal of Experimental Child Psychology* 78:3–24.

Needham, A., and R. Baillargeon. 2000. Infants' use of featural and experiential information in segregating and individuating objects: A reply to Xu, Carey, and Welch 1999. *Cognition* 74:255–284.

Nelson, K. 2005. Emerging levels of consciousness in early human development. In *The Missing Link in Cognition: Origins of Self-Reflective Consciousness*, ed. H. Terrace and J. Metcalfe. New York: Oxford University Press.

Neisser, U. 1991. Two perceptually given aspects of the self and their development. *Developmental Review* 11:197–209.

Newen, A., and A. Bartels. 2007. Animal minds and the possession of concepts. *Philosophical Psychology* 20:283–308.

Newen, A., and K. Vogeley. 2003. Self-representation: searching for a neural signature of self-consciousness. *Consciousness and Cognition* 12:529–543.

Ney, A. 2008. Reductionism. In *Internet Encyclopedia of Philosophy*, http://www.iep.utm.edu/red-ism.

Nichols, S., and S. Stich. 2003. *Mindreading*. New York: Oxford University Press.

Nickel, B. 2007. Against intentionalism. *Philosophical Studies* 136:279–304.

Nielsen, T. 1963. Volition: A new experimental approach. *Scandinavian Journal of Psychology* 4:225–230.

Nisbett, R., and T. Wilson. 1977. Telling more than we can know. *Psychological Review* 84:231–295.

Noë, A., ed. 2002. *Is the Visual World a Grand Illusion?* Special issue, *Journal of Consciousness Studies* 9 (5–6).

Noë, A. 2004. *Action in Perception*. Cambridge, MA: MIT Press.

Noë, A. 2006. Experience of the world in time. *Analysis* 66:26–32.

Noë, A. 2007. Inattentional blindness, change blindness, and consciousness. In *The Blackwell Companion to Consciousness*, ed. M. Velmans and S. Schneider. Malden, MA: Blackwell.

Oakes, L., and P. Bauer, eds. 2007. *Short- and Long-Term Memory in Infancy and Early Childhood*. Oxford: Oxford University Press.

O'Reilly, R., R. Busby, and R. Soto. 2003. Three forms of binding and their neural substrates: Alternatives to temporal synchrony. In *The Unity of Consciousness: Binding, Integration, and Dissociation*, ed. A. Cleeremans. Oxford: Oxford University Press.

Özgen, E., P. Sowden, P. Schyns, and C. Daoutis. 2005. Top-down attentional modulation of spatial frequency processing in scene perception. *Visual Cognition* 12:925–937.

Papineau, D. 1987. *Reality and Representation*. Oxford: Blackwell.

Papineau, D. 1998. Mind the gap. In *Philosophical Perspectives*, vol. 12, ed. J. Tomberlin. Atascadero, CA: Ridgeview.

Papineau, D. 2002. *Thinking about Consciousness*. Oxford: Oxford University Press.

Pascalis, O., S. DeSchonen, J. Morton, C. Deruelle, and M. Fabre-Grenet. 1995. Mother's face recognition by neonates: A replication and an extension. *Infant Behavior and Development* 18:79–85.

Pascual-Leone, A., and V. Walsh. 2001. Fast backprojections from the motion to the primary visual area necessary for visual awareness. *Science* 292:510–513.

Pauen, S. 2002. Evidence for knowledge-based category discrimination in infancy. *Child Development* 73:1016–1033.

Pautz, A. 2006. Sensory awareness is not a wide physical relation: An empirical argument against externalist intentionalism. *Noûs* 40:205–240.

Peacocke, C. 1989. Perceptual content. In *Themes from Kaplan*, ed. J. Almog, J. Perry, and H. Wettstein. Oxford: Oxford University Press.

Peacocke, C. 1992. *A Study of Concepts*. Cambridge, MA: MIT Press.

Peacocke, C. 2001a. Does perception have a nonconceptual content? *Journal of Philosophy* 98:239–264.

Peacocke, C. 2001b. Phenomenology and nonconceptual content. *Philosophy and Phenomenological Research* 62:609–615.

Penn, D., and D. Povinelli. 2007. On the lack of evidence that non-human animals possess anything remotely resembling a "theory of mind." *Philosophical Transactions of the Royal Society B* 362:731–744.

Perner, J., and Z. Dienes. 2003. Developmental aspects of consciousness: How much theory of mind do you need to be consciously aware? *Consciousness and Cognition* 12:63–82.

Perrett, R. 2003. Intentionality and self-awareness. *Ratio* 16:222–235.

Perry, J. 2001. *Knowledge, Possibility, and Consciousness*. Cambridge, MA: MIT Press.

Perry, E., H. Ashton, and A. Young, eds. 2002. *Neurochemistry of Consciousness*. Amsterdam: John Benjamins.

Phillips, A., H. Wellman, and E. Spelke. 2002. Infants' ability to connect gaze and emotional expression to intentional action. *Cognition* 85:53–78.

Piaget, J. 1954. *The Construction of Reality in the Child*. New York: Basic Books.

Picciuto, V. 2011. Phenomenal concepts and the nature of phenomenal consciousness. *Journal of Consciousness Studies* 18 (3–4):109–136.

Pitt, D. 2004. The phenomenology of cognition, or, what is it like to think that P? *Philosophy and Phenomenological Research* 69:1–36.

Place, U. T. 1956. Is consciousness a brain process? *British Journal of Psychology* 47:44–50.

Pollen, D. 1999. On the neural correlates of visual perception. *Cerebral Cortex* 9:4–19.

Pollen, D. 2003. Explicit neural representations, recursive neural networks and conscious visual perception. *Cerebral Cortex* 13:807–814.

Povinelli, D. 2000. *Folk Physics for Apes*. New York: Oxford University Press.

Povinelli, D., and J. Vonk. 2004. We don't need a microscope to explore the chimpanzee's mind. *Mind and Language* 19:1–28.

Povinelli, D., and J. Vonk. 2006. We don't need a microscope to explore the chimpanzee's mind. In *Rational Animals?*, ed. S. Hurley and M. Nudds. New York: Oxford University Press.

Premack, D., and G. Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1:515–526.

Prinz, J. 2000. A neurofunctional theory of visual consciousness. *Consciousness and Cognition* 9:243–259.

Prinz, J. 2002. *Furnishing the Mind: Concepts and Their Perceptual Basis*. Cambridge, MA: MIT Press.

Prinz, J. 2007. The intermediate level theory of consciousness. In *The Blackwell Companion to Consciousness*, ed. M. Velmans and S. Schneider. Malden, MA: Blackwell.

Proust, J. 2006. Rationality and metacognition in non-human animals. In *Rational Animals?* ed. S. Hurley and M. Nudds. New York: Oxford University Press.

Proust, J. 2009. The representational basis of brute metacognition: A proposal. In *The Philosophy of Animal Minds*, ed. R. Lurz. Cambridge: Cambridge University Press.

Putnam, H. 1975. The meaning of "meaning." In H. Putnam, *Mind, Language, and Reality: Philosophical Papers*, vol. 2. Cambridge: Cambridge University Press.

Pylyshyn, Z. 2007. *Things and Places: How the Mind Connects with the World*. Cambridge, MA: MIT Press.

Quine, W. V. O. 1951. Two dogmas of empiricism. *Philosophical Review* 60:20–43.

Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA: MIT Press.

Quinn, P. 2003. Concepts are not just for objects: categorization of spatial relation information by infants. In *Early Category and Concept Development*, ed. D. Rakison and L. Oakes. New York: Oxford University Press.

Raby, C., D. Alexis, A. Dickinson, and N. Clayton. 2007. Planning for the future by western scrub-jays. *Nature* 445:919–921.

Raby, C., and N. Clayton. 2009. Prospective cognition in animals. *Behavioural Processes* 80:314–324.

Raffman, D. 1995. On the persistence of phenomenology. In *Conscious Experience*, ed. T. Metzinger. Paderborn: Ferdinand Schöningh.

Raftopoulos, A. 2009a. *Cognition and Perception*. Cambridge, MA: MIT Press.

Raftopoulos, A. 2009b. Reference, perception, and attention. *Philosophical Studies* 144:339–360.

Raftopoulos, A., and V. Müller. 2006. The phenomenal content of experience. *Mind and Language* 21:187–219.

Rakison, D. 2006. Make the first move: How infants learn about self-propelled objects. *Developmental Psychology* 42:900–912.

Rakison, D. 2007. Is consciousness in its infancy in infancy? In *The Interplay between Consciousness and Concepts*, ed. R. Gennaro. Exeter: Imprint Academic.

Rakison, D., and L. Oakes, eds. 2003. *Early Category and Concept Development*. New York: Oxford University Press.

Ramachandran, V., and S. Blakeslee. 1998. *Phantoms in the Brain*. New York: Quill.

Rauschecker, J. 1998. Cortical processing of complex sounds. *Current Opinion in Neurobiology* 8:516–521.

Reber, A. 1967. Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior* 6:855–863.

Reber, A. 1993. *Implicit Learning and Tacit Knowledge*. New York: Oxford University Press.

Reber, P., D. Gitelman, T. Parrish, and M. Mesulam. 2003. Dissociating explicit and implicit category knowledge with fMRI. *Journal of Cognitive Neuroscience* 15:574–685.

Reed, J., L. Squire, A. Patalano, E. Smith, and J. Jonides. 1999. Learning about categories that are defined by object-like stimuli despite impaired declarative memory. *Behavioral Neuroscience* 113:411–419.

Revonsuo, A. 2006. *Inner Presence: Consciousness as a Biological Phenomenon*. Cambridge, MA: MIT Press.

Revonsuo, A. 2010. *Consciousness: The Science of Subjectivity*. New York: Psychology Press.

Rey, G. 1998. A narrow representationalist account of qualitative experience. *Philosophical Perspectives* 12:435–457.

Rhemtulla, M., and F. Xu. 2007. Sortal concepts and causal continuity: Comments on Rips, Blok, and Newman (2006). *Psychological Review* 114:1087–1095.

Riddoch, M., and G. Humphreys. 1987. A case of integrative visual agnosia. *Brain* 110:1431–1462.

Ridge, M. 2001. Taking solipsism seriously: Nonhuman animals and meta-cognitive theories of consciousness. *Philosophical Studies* 103:315–340.

Rips, L., S. Blok, and G. Newman. 2006. Tracing the identity of objects. *Psychological Review* 113:1–30.

Roberts, W. 2002. Are animals stuck in time? *Psychological Bulletin* 128:473–489.

Roberts, R. 2009. The sophistication of non-human emotion. In *The Philosophy of Animal Minds*, ed. R. Lurz. Cambridge: Cambridge University Press.

Robinson, W. 2004. *Understanding Phenomenal Consciousness*. New York: Cambridge University Press.

Rochat, P. 2001. *The Infant's World*. Cambridge, MA: Harvard University Press.

Rochat, P. 2003. Five levels of self-awareness as they unfold early in life. *Consciousness and Cognition* 12:717–731.

Rolls, E. 2004. A higher order syntactic thought (HOST) theory of consciousness. In *Higher-Order Theories of Consciousness: An Anthology*, ed. R. Gennaro. Amsterdam: John Benjamins.

Rosch, E., and C. Mervis. 1975. Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology* 7:573–605.

Rosenthal, D. M. 1986. Two concepts of consciousness. *Philosophical Studies* 49: 329–359.

Rosenthal, D. M. 1990. On being accessible to consciousness. *Behavioral and Brain Sciences* 13:621–622.

Rosenthal, D. M. 1991. The independence of consciousness and sensory quality. *Philosophical Issues* 1:15–36.

Rosenthal, D. M. 1993a. State consciousness and transitive consciousness. *Consciousness and Cognition* 2:355–363.

Rosenthal, D. M. 1993b. Thinking that one thinks. In *Consciousness: Psychological and Philosophical Essays*, ed. M. Davies and G. Humphreys. Oxford: Blackwell.

Rosenthal, D. M. 1997. A theory of consciousness. In *The Nature of Consciousness*, ed. N. Block, O. Flanagan, and G. Güzeldere. Cambridge, MA: MIT Press.

Rosenthal, D. M. 2000a. Consciousness and metacognition. In *Metarepresentation*, ed. D. Sperber. Oxford: Oxford University Press.

Rosenthal, D. M. 2000b. Consciousness, content, and metacognitive judgments. *Consciousness and Cognition* 9:203–214.

Rosenthal, D. M. 2000c. Metacognition and higher-order thoughts. *Consciousness and Cognition* 9:231–242.

Rosenthal, D. M. 2002a. Explaining consciousness. In *Philosophy of Mind: Classical and Contemporary Readings*, ed. D. Chalmers. New York: Oxford University Press.

Rosenthal, D. M. 2002b. How many kinds of consciousness? *Consciousness and Cognition* 11:653–665.

Rosenthal, D. M. 2003. Unity of consciousness and the self. *Proceedings of the Aristotelian Society* 103:325–352.

Rosenthal, D. M. 2004. Varieties of higher-order theory. In *Higher-Order Theories of Consciousness: An Anthology*, ed. R. Gennaro. Amsterdam: John Benjamins.

Rosenthal, D. M. 2005. *Consciousness and Mind*. New York: Oxford University Press.

Rosenthal, D. M. 2007. Phenomenological overflow and cognitive access. *Behavioral and Brain Sciences* 30:522–523.

Rosenthal, D. M. 2008. Consciousness and its function. *Neuropsychologia* 46:829–840.

Roskies, A. 2008. A new argument for nonconceptual content. *Philosophy and Phenomenological Research* 76:633–659.

Roskies, A. 2010. "That" response doesn't work: Against a demonstrative defense of conceptualism. *Noûs* 44:112–134.

Rovee-Collier, C., H. Hayne, and M. Colombo. 2001. *The Development of Implicit and Explicit Memory*. Amsterdam: John Benjamins.

Rowlands, M. 2007. Mysterianism. In *The Blackwell Companion to Consciousness*, ed. M. Velmans and S. Schneider. Malden, MA: Blackwell.

Rumelhart, D., and J. McClelland. 1986. *Parallel Distributed Processing*. Vols. 1 and 2. Cambridge, MA: MIT Press.

Rupert, R. 1999. The best bet theory of extension: First principle(s). *Mind and Language* 14:321–355.

Rupert, R. 2008. Causal theories of mental content. *Philosophy Compass* 3:353–380.

Ryder, D. 2004. SINBAD neurosemantics: A theory of mental representation. *Mind and Language* 19:211–240.

Sacks, O. 1987. *The Man Who Mistook His Wife for a Hat and Other Clinical Tales*. New York: Harper and Row.

Saidel, E. 2009. Attributing mental representations to animals. In *The Philosophy of Animal Minds*. ed. Robert Lurz. Cambridge: Cambridge University Press.

Samet, J. 1986. Troubles with Fodor's nativism. In *Midwest Studies in Philosophy*, vol. 10, ed. P. French, T. Uehling, and H. Wettstein. Minneapolis: University of Minnesota Press.

Samuels, R. 2002. Nativism in cognitive science. *Mind and Language* 17:233–265.

Samuels, R. 2007. Is innateness a confused concept? In *The Innate Mind: Foundations and the Future*, ed. P. Carruthers, S. Laurence, and S. Stich. New York: Oxford University Press.

Santos, L., A. Nissen, and J. Ferrugia. 2006. Rhesus monkeys, *Macaca mulatta*, know what others can and cannot hear. *Animal Behaviour* 71:1175–1181.

Sartre, J. 1956. *Being and Nothingness*. New York: Philosophical Library.

Scholl, B. 2007. Object persistence in philosophy and psychology. *Mind and Language* 22:563–591.

Schröder, J. 2001. Higher-order thought and naturalist accounts of consciousness. *Journal of Consciousness Studies* 8 (11):27–46.

Schwartz, B. 2005. Do nonhuman primates have episodic memory? In *The Missing Link in Cognition: Origins of Self-Reflective Consciousness*, ed. H. Terrace and J. Metcalfe. New York: Oxford University Press.

Schyns, P. 1998. Diagnostic recognition: Task constraints, object information, and their interactions. *Cognition* 67:147–179.

Schyns, P., L. Bonnar, and F. Gosselin. 2002. Show me the features! Understanding recognition from the use of visual information. *Psychological Science* 13:402–408.

Schyns, P., and A. Oliva. 1999. Dr. Angry and Mr. Smile: When categorization flexibly modifies the perception of faces in rapid visual presentations. *Cognition* 69:243–265.

Seager, W. 1999. *Theories of Consciousness*. New York: Routledge.

Seager, W. 2004. A cold look at HOT theory. In *Higher-Order Theories of Consciousness: An Anthology*, ed. R. Gennaro. Amsterdam: John Benjamins.

Seager, W., and D. Bourget. 2007. Representationalism about consciousness. In *The Blackwell Companion to Consciousness*, ed. M. Velmans and S. Schneider. Malden, MA: Blackwell.

Searle, J. 1992. *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Searle, J. 1995. Consciousness, the brain, and the connection principle: A reply. *Philosophy and Phenomenological Research* 55:217–232.

Searle, J. 2002. *Consciousness and Language*. Cambridge: Cambridge University Press.

Segal, G. 2000. *A Slim Book about Narrow Content*. Cambridge, MA: MIT Press.

Seger, C. 1994. Implicit learning. *Psychological Bulletin* 115:163–196.

Sellars, W. 1956. Empiricism and the philosophy of mind. In *Minnesota Studies in the Philosophy of Science*, ed. H. Feigl and M. Scriven. Minneapolis: Minnesota University Press.

Seth, A., and B. Baars. 2005. Neural Darwinism and consciousness. *Consciousness and Cognition* 14:140–168.

Seth, A., B. Baars, and D. Edelman. 2005. Criteria for consciousness in humans and other mammals. *Consciousness and Cognition* 14:119–139.

Shani, I. 2007. Consciousness and the first person. *Journal of Consciousness Studies* 14:57–91.

Shani, I. 2008. Against consciousness chauvinism. *Monist* 91:294–323.

Shea, N., and C. Heyes. 2010. Metamemory as evidence of animal consciousness: The type that does the trick. *Biology and Philosophy* 25:95–110.

Shear, J. 1997. *Explaining Consciousness: The Hard Problem*. Cambridge, MA: MIT Press.

Shoemaker, S. 1982. The inverted spectrum. *Journal of Philosophy* 79:357–381.

Shoemaker, S. 1994. Self-knowledge and "inner sense." *Philosophy and Phenomenological Research* 54:249–314.

Shoemaker, S. 2003. Consciousness and co-consciousness. In *The Unity of Consciousness: Binding, Integration, and Dissociation*, ed. A. Cleeremans. Oxford: Oxford University Press.

Shriver, A. 2006. Minding mammals. *Philosophical Psychology* 19:433–442.

Siegel, S. 2006. Which properties are represented in perception? In *Perceptual Experience*, ed. T. Gendler and J. Hawthorne. New York: Oxford University Press.

Siegel, S. 2009. The visual experience of causation. *Philosophical Quarterly* 59:519–540.

Siewart, C. 1998. *The Significance of Consciousness*. Princeton: Princeton University Press.

Siewert, C. 2008. Consciousness and intentionality. In *The Stanford Encyclopedia of Philosophy* (fall 2008 edition), ed. Edward N. Zalta, http://plato.stanford.edu/archives/fall2008/entries/consciousness-intentionality.

Simons, D. 2000. Current approaches to change blindness. *Visual Cognition* 7:1–15.

Simons, D., and C. Chabris. 1999. Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception* 28:1059–1074.

Simons, P. 1987. *Parts: A Study in Ontology*. Oxford: Clarendon Press.

Slater, A., and V. Morison. 1985. Shape constancy and slant perception at birth. *Perception* 14:337–344.

Slater, C. 1994. Discrimination without indication: Why Dretske can't lean on learning. *Mind and Language* 9:163–180.

Smart, J. J. C. 1959. Sensations and brain processes. *Journal of Philosophy* 68:141–156.

Smith, A. 2002. *The Problem of Perception*. Cambridge, MA: Harvard University Press.

Smith, D. W. 1986. The structure of (self-)consciousness. *Topoi* 5:149–156.

Smith, D. W. 1989. *The Circle of Acquaintance*. Dordrecht: Kluwer.

Smith, D. W. 2004. *Mind World: Essays in Phenomenology and Ontology*. Cambridge, MA: Cambridge University Press.

Smith, E. 2008. The case for implicit category learning. *Cognitive, Affective and Behavioral Neuroscience* 8:3–16.

Smith, E., and D. Medin. 1981. *Categories and Concepts*. Cambridge, MA: Harvard University Press.

Smith, J. D. 2005. Studies of uncertainty monitoring and metacognition in animals. In *The Missing Link in Cognition: Origins of Self-Reflective Consciousness*, ed. H. Terrace and J. Metcalfe. New York: Oxford University Press.

Smith, J. D., W. Shields, and D. Washburn. 2003. The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences* 26:317–373.

Smith, Q., and A. Jokic, eds. 2003. *Consciousness: New Philosophical Perspectives*. New York: Oxford University Press.

Sober, E. 1998. Morgan's canon. In *The Evolution of Mind*, ed. D. Cummins and C. Allen. Oxford: Oxford University Press.

Son, L., and N. Kornell. 2005. Metaconfidence judgments in rhesus macaques: Explicit versus implicit mechanisms. In *The Missing Link in Cognition: Origins of Self-Reflective Consciousness*, ed. H. Terrace and J. Metcalfe. New York: Oxford University Press.

Sosa, E., and M. Steup, eds. 2005. *Contemporary Debates in Epistemology*. Oxford: Blackwell.

Song, H., and R. Baillargeon. 2008. Infants' reasoning about others' false perceptions. *Developmental Psychology* 44:1789–1795.

Speaks, J. 2005. Is there a problem about nonconceptual content? *Philosophical Review* 114:359–398.

Spelke, E. 1990. Principles of object perception. *Cognitive Science* 14:29–56.

Spelke, E. 1998. Nativism, empiricism, and the origins of knowledge. *Infant Behavior and Development* 21:181–200.

Spelke, E., K. Breilinger, J. Macomber, and K. Jacobsen. 1992. Origins of knowledge. *Psychological Review* 99:605–632.

Spelke, E., and K. Kinzler. 2007. Core knowledge. *Developmental Science* 10:89–96.

Spelke, E., and G. Van de Walle. 1993. Perceiving and reasoning about objects: Insights from infants. In *Spatial Representation*, ed. N. Eilan, R. McCarthy, and W. Brewer. Oxford: Basil Blackwell.

Sperling, G. 1960. The information available in brief visual presentations. *Psychological Monographs* 74:1–29.

Stadler, M., and P. Frensch. 1998. *Handbook of Implicit Learning*. Thousand Oaks, CA: Sage.

Stampe, D. 1977. Toward a causal theory of linguistic representation. In *Midwest Studies in Philosophy*, vol. 2, *Studies in the Philosophy of Language*, ed. P. French, T. Uehling, and H. K. Wettstein. Minneapolis: University of Minnesota Press.

Stephens, G., and G. Graham. 2000. *When Self-Consciousness Breaks*. Cambridge, MA: MIT Press.

Stephens, G., and G. Graham. 2007. Philosophical psychopathology and self-consciousness. In *The Blackwell Companion to Consciousness*, ed. M. Velmans and S. Schneider. Malden, MA: Blackwell.

Sterelny, K. 1989. Fodor's nativism. *Philosophical Studies* 55:119–141.

Stich, S. 1978. Beliefs and subdoxastic states. *Philosophy of Science* 45:499–518.

Stoljar, D. 2004. The argument from diaphanousness. In *New Essays in the Philosophy of Language and Mind: Special Issue of the Canadian Journal of Philosophy*, vol. 30, ed. M. Ezcurdia, R. Stainton, and C. Viger. Calgary: University of Calgary Press.

Stoljar, D. 2009. Physicalism. In *The Stanford Encyclopedia of Philosophy* (fall 2009 edition), ed. Edward N. Zalta, http://plato.stanford.edu/archives/fall2009/entries/physicalism.

Strawson, P. 1966. *The Bounds of Sense*. New York: Methuen.

Strawson, G. 2004. Real intentionality. *Phenomenology and the Cognitive Sciences* 3:287–313.

Stubenberg, L. 1998. *Consciousness and Qualia*. Philadelphia: John Benjamins Publishers.

Stueber, K. 2006. *Rediscovering Empathy*. Cambridge, MA: MIT Press.

Suddendorf, T., and M. Corballis. 2007. The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and Brain Sciences* 30:299–313.

Surian, L., S. Caldi, and D. Sperber. 2007. Attribution of beliefs by 13-month-old infants. *Psychological Science* 18:580–586.

Sutton, J. 2004. Are concepts mental representations or abstracta? *Philosophy and Phenomenological Research* 68:89–108.

Terrace, H., and J. Metcalfe, eds. 2005. *The Missing Link in Cognition: Origins of Self-Reflective Consciousness*. New York: Oxford University Press.

Teuber, H. 1968. Alteration of perception and memory in man. In *Analysis of Behavioral Change*, ed. L. Weiskrantz. New York: Harper & Row.

Textor, M. 2006. Brentano (and some neo-Brentanians) on inner consciousness. *Dialectica* 60:411–432.

Thau, M. 2002. *Consciousness and Cognition*. Oxford: Oxford University Press.

Thomasson, A. 2000. After Brentano: A one-level theory of consciousness. *European Journal of Philosophy* 8:190–209.

Tomasello, M., J. Call, and B. Hare. 2003. Chimpanzees understand psychological states: The question is which ones and to what extent. *Trends in Cognitive Sciences* 7:153–156.

Tomasello, M., and J. Call. 2006. Do chimpanzees know what others see—or only what they are looking at? In *Rational Animals?* ed. S. Hurley and M. Nudds. New York: Oxford University Press.

Tong, F. 2003. Primary visual cortex and visual awareness. *Nature Reviews: Neuroscience* 4:219–229.

Tononi, G. 2004. An information integration theory of consciousness. *Biomedical Central Neuroscience* 5:1–22.

Toribio, J. 2007. Nonconceptual content. *Philosophy Compass* 2–3:445–460.

Treisman, A. 1993. The perception of features and objects. In *Attention: Selection, Awareness and Control. A Tribute to Donald Broadbent*, ed. A. Baddeley and L. Weiskrantz. Oxford: Clarendon Press.

Treisman, A. 2003. Consciousness and perceptual binding. In *The Unity of Consciousness: Binding, Integration, and Dissociation*, ed. A. Cleeremans. Oxford: Oxford University Press.

Trevarthen, C., and V. Reddy. 2007. Consciousness in infants. In *The Blackwell Companion to Consciousness*, ed. M. Velmans and S. Schneider. Malden, MA: Blackwell.

Tulving, E. 1983. *Elements of Episodic Memory*. Oxford: Oxford University Press.

Tulving, E. 1993. What is episodic memory? *Current Perspectives in Psychological Science* 2:67–70.

Tulving, E. 2005. Episodic memory and autonoesis: Uniquely human? In *The Missing Link in Cognition: Origins of Self-Reflective Consciousness*, ed. H. Terrace and J. Metcalfe. New York: Oxford University Press.

Tye, M. 1995. *Ten Problems of Consciousness*. Cambridge, MA: MIT Press.

Tye, M. 2000. *Consciousness, Color, and Content*. Cambridge, MA: MIT Press.

Tye, M. 2002. Representationalism and the transparency of experience. *Noûs* 36: 137–151.

Tye, M. 2003. *Consciousness and Persons*. Cambridge, MA: MIT Press.

Tye, M. 2006a. The thesis of nonconceptual content. *European Review of Philosophy* 6:7–30.

Tye, M. 2006b. Nonconceptual content, richness, and fineness of grain. In *Perceptual Experience*, ed. T. Gendler and J. Hawthorne. New York: Oxford University Press.

Tye, M. 2009a. Qualia. In *The Stanford Encyclopedia of Philosophy* (summer 2009 edition), ed. Edward N. Zalta, http://plato.stanford.edu/archives/sum2009/entries/qualia.

Tye, M. 2009b. *Consciousness Revisited: Materialism without Phenomenal Concepts*. Cambridge, MA: MIT Press.

Usher, M. 2001. A statistical referential theory of content: using information theory to account for misrepresentation. *Mind and Language* 16:311–334.

Van Gulick, R. 1985. Physicalism and the subjectivity of the mental. *Philosophical Topics* 13:51–70.

Van Gulick, R. 1993. Understanding the phenomenal mind: Are we all just armadillos? In *Consciousness: Psychological and Philosophical Essays*, ed. M. Davies and G. Humphreys. Oxford: Blackwell.

Van Gulick, R. 1995a. How should we understand the relation between intentionality and phenomenal consciousness? *Philosophical Perspectives* 9:271–289.

Van Gulick, R. 1995b. Why the connection argument doesn't work. *Philosophy and Phenomenological Research* 55:201–207.

Van Gulick, R. 1995c. What would count as explaining consciousness? In *Conscious Experience*, ed. T. Metzinger. Paderborn: Ferdinand Schöningh.

Van Gulick, R. 2000. Inward and upward: Reflection, introspection, and self-awareness. *Philosophical Topics* 28:275–305.

Van Gulick, R. 2004. Higher-order global states (HOGS): An alternative higher-order model of consciousness. In *Higher-Order Theories of Consciousness: An Anthology*, ed. R. Gennaro. Amsterdam: John Benjamins.

Van Gulick, R. 2006. Mirror mirror—is that all? In *Self-Representational Approaches to Consciousness*, ed. U. Kriegel and K. Williford. Cambridge, MA: MIT Press.

Varela, F., and E. Thompson. 2003. Neural synchrony and the unity of mind: a neurophenomenological perspective. In *The Unity of Consciousness: Binding, Integration, and Dissociation*, ed. A. Cleeremans. Oxford: Oxford University Press.

Varley, R. 1998. Aphasic language, aphasic thought: Propositional thought in an apropositional aphasic. In *Language and Thought: Interdisciplinary Themes*, ed. P. Carruthers and J. Boucher. Cambridge: Cambridge Yniversity Press.

Varzi, A. 2010. Mereology. In *The Stanford Encyclopedia of Philosophy* (spring 2010 edition), ed. Edward N. Zalta, http://plato.stanford.edu/archives/spr2010/entries/mereology.

Velmans, M., and S. Schneider, eds. 2007. *The Blackwell Companion to Consciousness*. Malden, MA: Blackwell.

Waskan, J. 2010. Connectionism. In *Internet Encyclopedia of Philosophy*, http://www.iep.utm.edu/connect.

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. 2011. Misrepresenting consciousness. *Philosophical Studies* 154:409–433.

Weiskopf, D. 2007. Concept empiricism and the vehicles of thought. In *The Interplay between Consciousness and Concepts*, ed. R. Gennaro. Exeter: Imprint Academic.

Weiskopf, D. 2008. The origins of concepts. *Philosophical Studies* 140:359–384.

Weiskopf, D. 2009. The plurality of concepts. *Synthese* 169:145–173.

Weiskrantz, L. 1986. *Blindsight*. Oxford: Clarendon.

Weiskrantz, L. 1997. *Consciousness Lost and Found*. New York: Oxford University Press.

Williford, K. 2006. The self-representational structure of consciousness. In *Self-Representational Approaches to Consciousness*, ed. U. Kriegel and K. Williford. Cambridge, MA: MIT Press.

Wimmer, H., and J. Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13:103–128.

Wright, E., ed. 2008. *The Case for Qualia*. Cambridge, MA: MIT Press.

Wright, W. 2003. McDowell, demonstrative concepts, and nonconceptual representational content. *Disputatio* 14:37–51.

Wright, W. 2005. Distracted drivers and unattended experience. *Synthese* 144:41–68.

Wynn, K. 1992. Children's acquisition of the number words and the counting system. *Cognitive Psychology* 24:220–251.

Xu, F. 2002. The role of language in acquiring object kind concepts in infancy. *Cognition* 85:223–250.

Xu, F. 2007. Sortal concepts, object individuation, and language. *Trends in Cognitive Sciences* 11:400–406.

Xu, F., and S. Carey. 1996. Infants' metaphysics: The case of numerical identity. *Cognitive Psychology* 30:111–153.

Xu, F., and S. Carey. 2000. The emergence of kind concepts: A rejoinder to Needham and Baillargeon. *Cognition* 74:285–301.

Xu, F., S. Carey, and N. Quint. 2004. The emergence of kind-based object individuation in infancy. *Cognitive Psychology* 49:155–190.

Xu, F., S. Carey, and J. Welch. 1999. Infants' ability to use object kind information for object individuation. *Cognition* 70:137–166.

Yablo, S. 1999. Concepts and consciousness. *Philosophy and Phenomenological Research* 59:455–463.

Young, A. 2003. Face recognition with and without awareness. In *The Unity of Consciousness: Binding, Integration, and Dissociation*, ed. A. Cleeremans. Oxford: Oxford University Press.

Zahavi, D. 1998. Brentano and Husserl on self-awareness. *Etudes Phénoménologiques* 27–28:127–169.

Zahavi, D., ed. 2000. *Exploring the Self*. Amsterdam: John Benjamins.

Zahavi, D. 2004a. Back to Brentano? *Journal of Consciousness Studies* 11 (10–11): 66–87.

Zahavi, D. 2004b. The embodied self-awareness of the infant: A challenge to the theory-theory of mind? In *The Structure and Development of Self-Consciousness*, ed. D. Zahavi, T. Grünbaum, and J. Parnas. Amsterdam: John Benjamins.

Zahavi, D. 2005. *Subjectivity and Selfhood*. Cambridge, MA: MIT Press.

Zahavi, D., T. Grünbaum, and J. Parnas, eds. 2004. *The Structure and Development of Self-Consciousness*. Amsterdam: John Benjamins.

Zeki, S. 2007. A theory of micro-consciousness. In *The Blackwell Companion to Consciousness*, ed. M. Velmans and S. Schneider. Malden, MA: Blackwell.

Zelazo, P., H. Gao, and R. Todd. 2007. The development of consciousness. In *The Cambridge Handbook of Consciousness*, ed. P. Zelazo, M. Moscovitch, and E. Thompson. Cambridge: Cambridge University Press.

Zelazo, P., M. Moscovitch, and E. Thompson, eds. 2007. *The Cambridge Handbook of Consciousness*. Cambridge: Cambridge University Press.

Zentall, T. 2005. Animals may not be stuck in time. *Learning and Motivation* 36: 208–225.

Index

Achromatopsia, 293–294
Acquaintance, 107–108
Agnosia, visual, 156–160, 179, 293–294
AIR theory of consciousness, 325n9
Allen, C., 144, 249, 251–252
Allen-Hermanson, S., 252–255
Ambiguous figures, 151–156, 179
Analytic/synthetic statements, 82–84
Animals, 23, 229–268
 and conceptualism, 262–268
 and I-thoughts, 237–248
 and memory, 239–242
 and mindreading, 244–247
 and pain, 232–237
Appearance-reality task, 260–261
A priori/a posteriori knowledge, 82–83
Armstrong, D., 49, 121
Ashby, F., 286–287
Asymmetric dependence theory, 35
Attention, 6, 52–53, 108–110, 215–218, 293. *See also* Awareness
 and consciousness, 169–172
 joint, 224
Autism, 257–262
Awareness, 6, 9
 focal (attentive), 109–114, 116–134, 161–163
 peripheral (inattentive), 109–114, 116–134, 161–163

Baars, B., 80, 100, 279–280, 285
Backward masking, 283
Baillargeon, R., 189, 191–192, 227
Baron-Cohen, S., 258, 260
Bayne, T., 292, 294–300
Beeckmans, J., 325n8
Behne, T., 224
Beliefs, 22–27, 226–228
Bennett, J., 188, 238, 242, 252
Bermúdez, J., 138–139, 144–145, 156, 180, 198, 256–257
Binding Problem, 91, 291–302
Birch, S., 227
Blindness, change/inattentional, 162–163, 170–171
Blindsight, 118
Block, N., 8, 21, 43, 122–123, 166–167, 276, 278, 282
Bloom, P., 205, 227
Botterell, A., 82
Boucher, J., 261
Brain, 270–291
Brentano, F., 15, 103, 106, 108, 111, 115
Brewer, B., 137, 174–176, 181–182, 215, 263–264
Brook, A., 107, 292, 296
Browne, D., 255
Bullier, J., 283
Buras, T., 107–108
Byrne, A., 42, 71–72, 165, 172–173, 267

Call, J., 255
Campbell, J., 217
Capgras delusion, 294

Carey, S., 189, 196, 203
Carruthers, P., 6, 16–17, 19, 21, 26, 38, 41–42, 44–49, 140, 225, 228–239, 245, 254–256, 258, 261
Caston, V., 103
Causality, 194–196, 221, 223, 275
Chalmers, D., 5, 7, 12–13, 18, 21, 75–77, 79–84, 86–87, 275–276, 292, 294–299
Chomsky, N., 186
Chuard, P., 136, 163, 166, 169, 173–175
Churchland, P., 68, 282, 286
Clark, Andy, 147
Clark, Austen, 65
Clayton, N., 240, 245
Cleeremans, A., 289–292
Cognitive overload argument, 46
Coliva, A., 181
Color, 137, 174–177
Conceivability arguments, 18–21, 80–83, 306n3
Concept, 30, 51, 68, 77–79
 acquisition of, 180–183, 185–219, 288–290
 artifact, 199, 204–205
 classical model of, 141
 coarse-grained/fine-grained, 173–183
 comparative, 176–180, 197–198
 demonstrative, 19, 174–176, 180–181, 197–198, 215–219, 231
 exemplar theory of, 142
 indexical, 19–20
 learning, 187–188, 212–213, 288–289
 mental, 223–228, 230, 243–251
 natural kind, 199, 202–203
 object, 191, 199
 phenomenal, 18–20
 possession, 140–144, 155, 168, 177–179, 191–198, 202, 204, 214–215, 218–219, 225, 247–251
 prototype theory of, 141–142
 proxytype theory of, 142
 recognitional, 18–20
 reidentification condition of, 175–179
 sortal, 191–192
 testing methods for infants, 189–190
 theory-theory of, 142, 198 (*see also* Mindreading)
Conceptualism, 135–183
 and animals, 262–268
 and concept acquisition, 210–218
 definition of, 135–137
 and fineness of grain, 173–183
 and modality, 150–151
 and the richness of experience, 161–173
Conceptual pluralism, 142
Confabulation, 68–69, 97
Connectionism, 288–291
Connection Principle (CP), 21–25, 146–147, 234
Consciousness
 access, 8
 animal, 229–268
 definition of, 5–9
 hard problem of, 75–88
 infant, 219–228
 and intentionality, 11–12, 21–28
 intrinsic/extrinsic, 57–58, 72
 paradox, 2
 phenomenal, 7–9, 41
 phenomenal character, 7, 39 124–125
 the real hard problem of, 185–186
 state/creature, 5, 276
 subjective/qualitative character, 7, 64–65, 124–125, 151
 transitive/intransitive, 6, 72
 unity of, 161, 291–302
Content, mental
 analog/digital, 140, 162–163
 causal theory of, 34–36, 107, 200–203, 217
 conceptual, 135–137 (*see also* Conceptualism)
 Fregean/Russellian, 13, 36–37, 44, 149
 narrow/wide, 12–13, 26, 37–38, 43–44

nonconceptual, 40–41, 138–140, 284
subpersonal, 145–147, 217, 219
Continuity argument, 262–268
Convergence zones, 287–288, 297
Core knowledge, 189–205, 212–213
Cortex. *See* Brain
Cowey, A., 279
Cowie, F., 204
Crick, F., 74, 77, 273, 275
Cross-order integration (COI) theory, 277–278

Dainton, B., 291, 298–299
Damasio, A., 281, 287
De Gardelle, V., 167
Dehaene, S., 280, 284
Del Cul, A., 280
Dennett, D., 80, 162, 300
Dental fear, 69
Dere, E., 241
Descartes, R., 29, 57, 72–73, 237, 250, 299
Disjunction problem, 34–36, 200–202
Dispositional HOT theory. *See* Dual-content theory
Dretske, F., 34, 93, 140, 161, 165, 170, 238
Droege, P., 100–101
Dual-content theory, 45–49

Ebbinghaus illusion (Titchener circles), 147
Edelman, G., 68, 272, 274, 282–283
Eichenbaum, H., 241
Eilan, N., 224
Emery, N., 245
Essentialism, psychological, 203
Evans, G., 139, 142, 174, 247
Evolution, 46–47, 126–127
Experience, 6. *See also* Consciousness
Explanation, 15–17, 42, 84–87, 105–108, 180, 255
Explanatory gap, 17–20

False-belief task, 226–228, 260–261, 322–323n13
Farah, M., 156–157, 159–160
Farrant, A., 261
Feature integration theory (FIT), 292–293
Feedback loops/reentrant feedback, 68, 98–99, 272, 282–288
Feeling of knowing, 70
Feinberg, T., 99–100, 285
FINSTs (fingers of instantiation), 145–146, 217, 219
First-order representationalism (FOR), 3, 13–14, 39–45, 237
Fitzpatrick, S., 254–255
Fivaz-Depeursinge, E., 223
Flohr, H., 274
Flombaum, J., 244–245
Fodor, J., 17, 34–35, 141–142, 146, 187–189, 199, 207, 212, 290
Ford, J., 127–129
Frege, G., 13, 36–37
Frith, C., 258–260, 262

Gelman, S., 203
Generality constraint, 142–143, 248–252
Georgalis, N., 36
Goldman, A., 71–72, 225, 322n12
Goldstone, R., 137
Goodale, M., 22, 42, 147
Gopnik, A., 198, 219–220
Grandin, T., 260
Gunther, Y., 135–136

Hampton, R., 241, 243
Happé, F., 258–260, 262
Hardcastle, V., 17, 32–33
Hard problem of consciousness, 75–88
Harman, G., 13
Hassin, R., 208
Hauser, M., 144, 267
Heck, R., 139, 180–181

Hellie, B., 107
Higher-order global states (HOGS), 15, 312–313n23
Higher-order perception (HOP) theory, 4, 14, 45, 49–53, 120–121, 231
Higher-order thought (HOT), 1, 14, 20, 30, 50–51
 targetless, 68–70
Higher-order thought (HOT) theory, 1–5, 14, 20, 28–33, 31 (fig.), 36–37, 43, 49, 57–58, 104–105, 195
 and animals, 229–247, 251, 257, 260–261, 278–279
 and autism, 257–262
 and brain, 276–282
 and conceptualism, 147–161, 165, 172–173, 178–180
 dispositional/actualist, 45–49
 and the hard problem, 75–88
 and infants, 220–228, 278–282
 noninferentiality condition, 32
 and regress/circularity, 30
 and the unity of consciousness, 294–302
Hill, C., 50–51, 69
Hillier, A., 258
Homomorphism theory, 65–66, 151
Horgan, T., 25–26
Hossack, K., 315–316n10
HOT-CON argument, 148
HOT Unity thesis, 298
Hume, D., 300
Humphreys, G., 159, 293–294
Hurley, S., 326n14

Identity theory, 17
Implicit Learning Theory, the (TILT), 206–210
Indeterminacy, 24
Infants, 189–198, 202–205, 208–212, 215, 219–228
Innate/nativism, 186–199, 211–215, 224–225
 core, 189–205, 212–213
 radical, 186–189
Intentionality, 11–12, 21–28, 34–36
 phenomenal, 25–27
Introspection, 30–32, 46–47, 51–53, 58, 69, 96, 259–260
 fallibility of, 53, 96–97, 110–113, 119–127, 131–132, 221, 250, 278–279, 300
Inverted Earth, 43
Inverted qualia, 43–44
I-thoughts, 220–223, 237–239, 248, 257–258. *See also* Metacognition

Jackson, F., 306n3
Jacoby, L., 32–33
James, W., 194
Janzen, G., 96–97, 108
Jeannerod, M., 220–221, 223
Jehle, D., 48
Jiang, Y., 171
Johnson, S., 224

Kant, I., 51, 77–79, 83, 89, 98, 120, 136, 149, 189, 193, 197, 242, 291, 300–301
Kantian categories, 189
Kastner, S., 137
Kelly, S., 139, 174–177
Kemeny, J., 16
Kéri, S., 287
Kihlstrom, J., 206–208, 214
Kind, A., 7, 40, 44, 123
Kinsbourne, M., 242, 244
Knowlton, B., 207
Koch, C., 77, 172, 273, 275–276, 279
Koivisto, M., 163
Koriat, A., 70, 238
Kornell, N., 243–244
Kosslyn, S., 285–286
Kouider, S., 167, 284
Kriegel, U., 7, 15, 36, 41–42, 48, 91–92, 95, 103–105, 107–108, 111–115, 117, 124–127, 130–134, 277–278
Kripke, S., 18, 82

Lamme, V., 100, 283
Language of thought, 141
Laurence, S., 187, 199, 202–203
Learning, implicit, 206–214, 225
Led Zeppelin, 118
Leekam, S., 258
Legrand, D., 220
Leibniz, G., 188, 265–266
Levine, J., 17, 58–65, 99, 148, 158, 160
Loar, B., 18–19, 123
Locke, J., 14, 188
Loewer, B., 64
Long-distance driver, 121–122
Lurz, R., 235, 238, 257, 309n9
Lycan, W., 12, 14, 28–29, 44, 49–53, 71, 76, 121
Lyyra, P., 257

Machery, E., 142
Mack, A., 162, 171
MacPherson, F., 152–155
Malach, R., 279
Mandler, J., 192–193, 203, 205, 209–210
Mareschal, D., 290
Margolis, E., 187, 199, 202–203
Marr, D., 145–146
Martin, M., 165, 167–168, 170
Materialism, 17–21, 81
Matey, J., 310n6
Matthen, M., 217
McClelland, J., 288
McDowell, J., 136–137, 174–175, 215, 264–266
McGinn, C., 17–18, 88, 120, 306n3
Medin, D., 203
Meltzoff, A., 198, 222
Memory, 128, 164–172, 175, 208–209, 221–222, 239–242, 261–262
Mental concepts, 223–228, 230, 243–251. *See also* Mindreading; Metacognition
Mental content. *See* Content, mental
Mental paint, 123
Mental stain, 123
Mental state attitude, 89–91
Mental state vehicle, 89–91
Mereology (parts/whole), 88–89, 91–95, 100–102, 114–116, 282–288, 296–297
Metacognition, 238, 243, 258, 260, 322n12
Metaconfidence judgments, 243–244
Metapsychological thought (MET), 55–59. *See also* Wide intrinsicality view (WIV)
Michotte, A., 195
Milner, D., 22, 42, 147
Mindreading, 225–226, 244–247, 260, 322n12
Misrepresentation, problem of, 49, 58, 59–70, 96, 99–100, 158–160, 179–180, 314n31
Mole, C., 170–171
Montminy, M., 252–253
Moore, G., 13
Morgan's Canon, 252–256
Morin, A., 220–221
Motion, 192–193
Müller-Lyer illusion, 154
Multiple realizability, 17
Myth of the given, 136

Nagel, T., 6
Natsoulas, T., 115
Neander, K., 49, 58–59, 63–64,
Necessary truths, 82–83, 86
Needham, A., 192
Neisser, U., 196
Nelson, K., 242
Nested hierarchy theory of consciousness (NHTC), 99–100, 285
Neural correlates of concepts, 286–287
Neural correlates of consciousness (NCCs), 269–270, 273–276, 301
Neurons, 68, 91, 98–100, 270–291
Newen, A., 280–281
Ney, A., 16

Nichols, S., 225, 258–259, 322n12
Nielsen, T., 221
Noë, A., 162, 198, 214, 268
Number, 196

Oakes, L., 190
Object, 158, 192–195
O'Reilly, R., 293

PANIC theory, 39–42, 140
Parallel distributed processing (PDP), 288
Parsimony. *See* Morgan's Canon
Pascual-Leone, A., 279
Past relative frequencies (PRFs), 35, 201, 287–288
Peacocke, C., 139, 141, 182, 262–263, 318n9
Perner, J., 226
Perrett, R., 109
Phenomenal intentionality, 25–27
Piaget, J., 191
Pollen, D., 100, 279, 283
Povinelli, D., 232, 245, 254
Prefrontal cortex (PFC), 277–282
Prinz, J., 35–36, 142, 186, 200–202, 287, 325n9
Priority argument, 180–183
Prosopagnosia, 160, 294
Psychopathologies, 127–129, 293–294, 297
Pure self-referentialism (PSR), 103–116, 105 (fig.), 123–124
Putnam, H., 12, 37
Pylyshyn, Z., 145–146, 216–217, 290

Qualia, 7–8, 30, 64–65, 78
Qualitative states, 7–8, 65, 92–93
Quine, W. V. O., 24, 84
Quinn, P., 192

Raby, C., 241
Raftopoulos, A., 145, 217–218
Rakison, D., 190, 210
Reber, A., 206, 208
Reduction, 15–21, 24, 106–107
Reed, J., 207
Reference, 12–13, 20, 38, 145
Reflection. *See* Introspection
Representation, 3, 9, 11–12, 34–36, 39, 44
Representationalism, 11–12, 117
 first-order, 13–14, 39–45, 237
 higher-order (*see* Higher-order thought theory)
 strong/weak, 12
 wide/narrow, 12–13, 43–44
Representation-as, 44, 148
Revonsuo, A., 5, 163, 269, 280
Riddoch, M., 159
Ridge, M., 246–247, 260
Rochat, P., 193, 196–197, 220–221, 226
Rock, I., 162, 171
Rock, problem of the, 58, 70–75
Rolls, E., 47, 280
Rosenthal, D., 1, 5–6, 14–15, 20, 28–32, 47, 57–60, 64–66, 68–70, 72–73, 76, 78, 89–90, 92–93, 97–98, 110, 148–149, 151, 172–173, 176, 278, 299–300
Roskies, A., 211–219
Rovee-Collier, C., 222
Rumelhart, D., 288
Rupert, R., 35–36, 200–201, 287–288
Russell, B., 13, 36–37

Samuels, R., 186
Santos, L., 244–245, 267
Schröder, J., 86–87, 101–102
Schwartz, B., 241
Schyns, P., 156
Seager, W., 238, 247–248
Searle, J., 21–25, 146, 234
Seeing-as, 68, 148, 151–153, 181–183
Self, 196–197, 220, 295, 299–301

Self-concepts, 196–197, 220–223, 241–244. *See also* I-thoughts
Self-consciousness, 111–112, 258–259. *See also* Introspection
bodily, 29, 196–197, 220, 242, 256
Self-representationalism, 15, 29, 103–134, 277
Sellars, W., 136, 151
Sensorimotor theory of consciousness, 326n14
Seth, A., 280
Shani, I., 22–23
Shields, W., 243
Shoemaker, S., 43, 292, 298
Siegel, S., 136, 194–196
Siewart, C., 307n11
Simons, D., 162–163
Smith, A., 156–157, 181–182, 214
Smith, D. W., 113–115, 125, 127–129, 131
Smith, E., 207
Smith, J., 243
Sober, E., 253
Son, L., 243–244
Song, H., 227
Space, 119–120, 192–193, 197
Speaks, J., 167, 175, 263, 267–268
Speckled hen, 168–169
Spelke, E., 189, 191–192, 194–195
Sperling, G., 164
Sperling experiment, 164, 166–167
Sterelny, K., 188
Stich, S., 24, 225, 258–259, 322n12
Strawson, P., 79
Stubenberg, L., 58, 70–71, 73–74, 76
Subliminal priming, 32–33
Subliminal processing, 284
Substance. *See* Object
Sustaining mechanism, 202–204

Teuber, H., 156
Theory-theory/theory of mind, 225–228, 232, 258, 281, 322n12
Thomasson, A., 96
Thompson, E., 326n14
Tienson, J., 26
TILT (The Implicit Learning Theory), 206–210
Time, 119–120, 193–194, 197
Tip of the tongue phenomenon, 70
Tomasello, M., 228, 255
Tong, F., 279
Tononi, G., 68, 272, 274, 282–283
Toribio, G., 135, 138, 142, 248
Transitivity Principle (TP), 28–29, 32, 47, 98, 113, 133
Transparency of experience, 13–14, 39–40, 116, 122–123, 132
Treisman, A., 292–293
Tulving, E., 221, 239–241
Twin Earth, 12, 37, 44
Two-system theory of vision, 22, 42, 147
Tye, M., 7, 13, 39–43, 138–140, 143–144, 153, 163–166, 168–169, 237, 296

Uncertainty monitoring, 243

Van Gulick, R., 15, 23, 49, 74, 83–86, 92, 312–313n23
Varela, F., 326n14
Varzi, A., 94
Vogeley, K., 280–281

Weisberg, J., 97–100, 314n31
Weiskopf, D., 142, 287
"What it's like," 6, 76, 78
Wide intrinsicality view (WIV), 4–5, 15, 56 (fig.), 55–59, 63–64, 88–102, 104–115, 282–291
Wimmer, H., 226
Wright, W., 121, 180
Wynn, K., 196

Xu, F., 204

Young, A., 294

Zahavi, D., 109, 223, 226, 228, 260–261
Zeki, S., 279
Zero-crossings, 145, 219
Zombies, 18–19, 80–81