REPRESENTATION RECONSIDERED

Cognitive representation is the single most important explanatory notion in the sciences of the mind and has served as the corner-stone for the so-called "cognitive revolution." This book critically examines the ways in which philosophers and cognitive scientists appeal to representations in their theories, and argues that there is considerable confusion about the nature of representational states. This has led to an excessive over-application of the notion – especially in many of the newer theories in computational neuroscience. *Representation Reconsidered* shows how psychological research is actually moving in a non-representational direction, revealing a radical, though largely unnoticed, shift in our basic understanding of how the mind works.

WILLIAM M. RAMSEY is Associate Professor in the Department of Philosophy, University of Notre Dame.

REPRESENTATION RECONSIDERED

WILLIAM M. RAMSEY
University of Notre Dame

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521859875

First published 2007

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

ISBN 978-0-521-85987-5 hardback

Contents

Figures

Preface

It has become almost a cliché to say that the most important explanatory posit today in cognitive research is the concept of representation. Like most clichés, it also happens to be true. Since the collapse of behaviorism in the 1950s, there has been no single theoretical construct that has played such a central role in the scientific disciplines of cognitive psychology, social psychology, linguistics, artificial intelligence, and the cognitive neurosciences. Of course, there have been many different types of representational theories. But all share the core assumption that mental processes involve content-bearing internal states and that a correct accounting of those processes must invoke structures that serve to stand for something else. The notion of mental representation is *the* corner-stone of what often gets referred to in Kuhnian terms as the "cognitive revolution" in psychology. But mental representation hasn't been important just to psychologists. Accompanying this trend in the sciences has been a corresponding focus on mental representation in the philosophy of mind. Much of this attention has focused upon the nature of commonsense notions of mental representation, like belief and desire, and how these can be part of a physical brain. More specifically, the central question has focused on the representational nature of beliefs – the fact that they have meaning and are essentially *about* various states of affairs.

Yet despite all of this attention (or perhaps because of it), there is nothing even remotely like a consensus on the nature of mental representation. Quite the contrary, the current state of affairs is perhaps best described as one of disarray and uncertainty. There are disagreements about how we should think about mental representation, about why representations are important for psychological and neurological processes, about what they are supposed to do in a physical system, about how they get their intentional content, and even about whether or not they actually exist. Part of this chaos is due to recent theoretical trends in cognitive science. The central explanatory framework behind a great deal of cognitive

research has traditionally been the classical computational theory of cognition. This framework regards the mind as a computational system with discrete internal symbols serving as representational states. However, over the past twenty years there have been dramatic departures from the classical computational framework, particularly with the emergence of theories in the cognitive neurosciences and connectionist modeling. These newer approaches to cognitive theorizing invoke radically different notions of cognitive representation; hence, they have generated considerable disagreement about how representation should be understood.

Still, debates over representation are not simply due to the existence of different cognitive theories and models. Often, the nature of representation *within* these different frameworks is unclear and disputed. One might expect some assistance on these matters from philosophers of psychology, especially given the amount of philosophical work recently focusing upon representation. Yet up to this point, it is far from obvious that philosophical work on representation has helped to ameliorate the situation in cognitive science. Philosophical work on representation has been a predominantly *a priori* enterprise, where intuitions about meaning are analyzed without special concern for the nuances of the different notions of representation that appear in scientific theories. While abstract questions about the nature of content are important, esoteric discussions about hypothetical scenarios, like the beliefs of Twin-Earthlings or spontaneously generated "swamp-men," have failed to be of much use to non-philosophers in the scientific community. Moreover, because of a preoccupation with the nature of content, philosophers have neglected other issues associated with cognitive representation that are more pressing to researchers. Of these other issues, perhaps the most important is explaining what it is for a neurological (or computational) state actually to *function as a representation* in a biological or computational system. Despite the importance of this issue to empirical investigators, the actual role representations are supposed to play, *qua representations*, is something that has received insufficient attention from philosophers.

My own interest in these matters began as a graduate student in the mid-1980s, with a front row seat on the exciting development of connectionist modeling taking place at the University of California, San Diego. A great deal of buzz was generated by the radically different picture of representation that accompanied connectionist models, especially their distributed and non-linguistic form. Yet every time I tried to get a clearer sense of just how, exactly, the internal nodes or connections were supposed to function as representational states, I failed to receive a satisfactory answer. Often my

queries would be met with a shrug and reply of "what else could they be doing?" It seemed the default assumption was that these hypothetical internal structures must be representations and that the burden of proof was upon anyone who wished to deny it. I first expressed my concerns about the explanatory value of connectionist representations much later, in a paper published in *Mind and Language* (Ramsey, 1997). At the time, William Bechtel correctly noted that my arguments, if they worked, would challenge not only the notions of representation associated with connectionism, but also the representational posits associated with a much wider range of theories. Although Bechtel intended this point as a problem with my view, I saw it as revealing a serious problem with the way people were thinking about representation within the broader cognitive science community.

Since that time, my skepticism about popular conceptions of representation has only grown, though not entirely across the board. I have also come to appreciate how some notions of representation actually do succeed in addressing my worries about representational function. To be sure, these notions of representation have their problems as well. But as the saying goes, there are problems and then there are *problems*. My belief is that some of the notions of representation we find in cognitive research need a little fixing up here and there, whereas other notions currently in vogue are hopeless non-starters. As it happens, the notions of representation that I think are promising are generally associated with the classical computational theory of cognition, whereas the notions I think are non-starters have been associated with the newer, connectionist and neurologically-based theories. Spelling all this out is one of the main goals of this book. The central question my analysis will ask is this: "Do the states characterized as representation in explanatory framework X actually serve as representations, given the processes and mechanisms put forth?" The answer I'm going to offer is, by and large, "yes" for the classical approach, and "no" for the newer accounts. When we look carefully at the way the classical framework explains cognitive processes, we find that talk of representation is justified, though this justification has been clouded in the past by misguided analyses. However, when we look at the explanatory strategies provided by the newer accounts, we find something very different. Although neuroscientific and connectionist theories characterize states and structures as inner representations, there is, on closer inspection, no compelling basis for this characterization.

It might be assumed that such an assessment would lead to an endorsement of the classical framework over the newer accounts. But that would

follow only if we presume that psychological theories absolutely must invoke representational states in their explanations of cognitive capacities. I think it is an open empirical question whether or not the brain actually uses representational states in various psychological processes. Most of the theories I criticize here still might prove workable, once the conceptual confusions about representation are cleared away. What my analysis does reveal, however, is that something very interesting is taking place in cognitive science. When new scientific theories are offered as alternatives to more established views, proponents of the new perspective are sometimes reluctant to abandon the familiar notions of the older framework, even when those posits have no real explanatory role in the new accounts. When this happens, the old notions may be re-worked as theorists contrive to fit them into an explanatory framework for which they are ill-suited. One of the central themes of this book is that something very much like this is currently taking place in cognitive science. My claim is that the representational perspective, while appropriate for classical computational cognitive science, has been carried over and assigned to new explanatory frameworks to which it doesn't actually apply. Although investigators who reject the classical framework continue to talk about internal representations, the models and theories many of them propose neither employ, nor need to employ, structures that are actually playing a representational role. I will argue that cognitive research is increasingly moving away from the representational paradigm, although this is hidden by misconceptions about what it means for something to serve as a representational state.

Thus, my primary objective is to establish both a positive and a negative thesis. The positive position is that, contrary to claims made by critics of conventional computationalism, the classical framework does indeed posit robust and explanatorily valuable notions of inner representation. To see this, we need to abandon what I call the "Standard Interpretation" of computational symbols as belief-like states, and instead view them as representations in a more technical sense. Computational explanation often appeals to mental models or simulations to account for how we perform various cognitive tasks. Computational symbols serve as elements of such models, and, as such, must *stand in for* (i.e., represent) elements or aspects of that which is being modeled. This is one way in which the classical picture employs a notion of representation that is doing real explanatory work. My negative claim is that the notions of representation invoked by many non-classical accounts of cognition do not have this sort of explanatory value. Structures that are described as representations are actually playing a functional role that, on closer inspection, turns out to

have little to with do anything recognizably representational in nature. For example, proposed structures are often characterized as representations because they faithfully respond to specific stimuli, and in turn causally influence other states and processes. My claim will be that this is not a representational role, and that these posits are better described as relay circuits or causal mediators.

In arguing for both the positive and negative theses, I will appeal to what I call the "job-description challenge." This is the challenge of explaining how a physical state actually fulfills the role of representing in physical or computational process – accounting for the way something actually *serves* as a representation in a cognitive system. In the philosophy of psychology, the emphasis upon content has led many to assume that a theory of content provides a theory of representation. But an account of content is only one part of the story. The question of how a physical structure comes to function as a representation is clearly different from (though related to) the question of how something that is presumed to function as a representation comes to have the intentional content it does. I claim that when we take the former question seriously, we can see that, by and large, classical computational representations meet the job-description challenge, but the notions of representation in the newer theories do not.

The analysis I will offer here is inspired by Robert Cummins's suggestion that the philosophy of psychology (and the philosophy of representation in particular) should primarily be an enterprise in the philosophy of science. Just as philosophers of physics might look at the explanatory role of the posits of quantum physics, or a philosopher of biology might look at different conceptions of genes, my agenda is to critically examine the different ways cognitive scientists appeal to notions of representation in their explanations of cognition. I believe such an assessment reveals that cognitive science has taken a dramatic anti-representational turn that has gone unnoticed because of various mis-characterizations of the posits of the newer theories. Cognitive theories are generally described as distinct from behaviorist accounts because they invoke inner representation. However, if many current cognitive theories are, as I argue, not actually representational theories, then we need to reconsider the scope of the so-called "cognitive revolution" and the degree to which modern cognitivism is really so different from certain forms of behaviorism. Moreover, a non-representational psychology would have important implications for our commonsense conception of the mind – our so-called "folk psychology." Since commonsense psychology is deeply committed to mental representations in the form of beliefs and other propositional attitudes, this

non-representational reorientation of cognitive science points in the direction of eliminative materialism – the radical thesis that beliefs don't actually exist. Eliminativism would bring about a cataclysmic shift in our understanding not just of psychological processes, but in our overall conception of ourselves. Thus, the developments that I will try to illuminate here are of enormous significance, despite having gone unnoticed by most cognitive scientists and philosophers of psychology.

To show all this, the book will have the following structure. In the first chapter, I introduce some of the issues and concerns that will take center stage in the subsequent chapters. After explaining the central goals of the book, I look at two families of representational concepts – one mental, the other non-mental – to get a preliminary handle on what it might mean to invoke representations as explanatory posits in cognitive science. I argue that our commonsense understanding of representation constrains what can be treated as a representation and presents various challenges for any scientific account of the mind that claims to be representational in nature. I also introduce the job description challenge and argue that theories that invoke representations carry the burden of demonstrating just how the proposed structure is supposed to serve as a representation in a physical system. Moreover, I argue this must be done in such a way that avoids making the notion of representation completely uninteresting and divorced from our ordinary understanding of what a representation actually is.

The goal of the second chapter is to present what I take to be a popular set of assumptions and tacit attitudes about the explanatory role of representation in the classical computational theory of the mind. I'll suggest that these assumptions and attitudes collectively give rise to an outlook on representation that amounts to a sort of merger between classical computational theory and folk psychology. This has led to a way of thinking about computational representations that suggests their primary explanatory function is to provide a scientific home for folk notions of mental representations like belief. I call this the "Standard Interpretation" of classical computationalism. After spelling out what I think the Standard Interpretation involves, I'll try to show that it leads us down a path where, despite various claims to the contrary, we wind up wondering whether the symbols of classical models should be viewed as representations at all. This path has been illuminated by two important skeptics of classical AI, John Searle and Stephen Stich. Searle and Stich both exploit the alleged link between classicalism and folk psychology to challenge the claim that the classical framework can or should appeal to inner

representations. I'll present Searle's and Stich's criticism of representationalism and examine the ways defenders of the Standard Interpretation have responded. In the final analysis, I'll argue the Standard Interpretation leaves in doubt the representational nature of computational states.

In the third chapter, I reject the Standard Interpretation and provide what I believe is the proper analysis of representation in the classical computational theory. Picking up on themes suggested by prior writers (such as John Haugeland and Robert Cummins), I argue that there are two related notions playing valuable explanatory roles, and that neither notion is based upon commonsense psychology. One notion pertains to the classical computational strategy of invoking inner computational operations to explain broader cognitive capacities. I argue that these inner subcomputations require inputs and outputs that must be representational in nature. The second notion, designated as "S-representation," pertains to data structures that in classical explanations serve as elements of a model or simulation. That is, according to many theories associated with the classical framework, the brain solves various cognitive problems by constructing a model of some target domain and, in so doing, employs symbols that serve to represent aspects of that domain. After providing a sketch of each notion, I consider two popular criticisms against them and argue that both criticisms can be handled by paying close attention to the way these notions are actually invoked in accounts of cognition. Finally, I address a number of side issues associated with these notions, such as their explanatory connection to computational rules and the question of whether they would vindicate the posits of folk psychology.

The fourth chapter begins the negative phase of the book and is devoted to exploring what I call the "receptor" notion of representation that appears in a wide range of theories in cognitive neuroscience and connectionist modeling. This style of representation often borrows from Shannon and Weaver's theory of information, and rests on the idea that neural or connectionist states represent certain stimuli because of a co-variance or nomic dependency relation with those stimuli. The work of Fred Dretske provides what is perhaps the clearest and most sophisticated defense of the explanatory value of this family of representational notions. However, despite Dretske's impressive support for this type of representation, I argue that the notion is too weak to have any real explanatory value. What gets characterized as a representation in this mold is often playing a functional role more akin to a non-representational relay circuit or simple causal mediator. In these cases, any talk of "information carrying" or representational content could be dropped altogether without any real

explanatory loss. I look closely at the arguments presented by Dretske and suggest that his account of representation is inadequate because it fails to meet the job description challenge.

The fifth chapter looks at a somewhat scattered family of representational notions found in various accounts of neurological processes, artificial intelligence and in various connectionist networks. Here the basic idea is that the functional architecture of a system plays a representational role largely because it is causally relevant to the production of various types of output. I characterize this as the "tacit" notion of representation since there is typically no one-to-one mapping between cognitive structures and individually represented items. The functional architecture of a system is said to encode information holistically, and this is thought to serves as the system's "know-how." After explaining the core features associated with this family of representational notions, I offer a critical evaluation and argue that, like the receptor notion, it fails to meet the job description challenge. Once again, representation is confused with something else; in this case, with the dispositional properties of the underlying architecture. Since there is no real motivation for treating these sorts of structures as representations, I defend the position that we should stop thinking of them in this way.

The sixth and final chapter addresses three important topics related to my analysis. First, to solidify my earlier claims, I offer a more direct comparison between the receptor and S-representational notions in the form of imaginary, quasi-robotic systems attempting to navigate a track. My aim here is to make clearer just how and why the receptor notion runs into trouble, while the S-representation notion is better suited for psychological theorizing. Second, in recent years, pockets of anti-representationalism have developed in various areas such as robotics research and Dynamic Systems Theory, and defenders of representationalism have offered a number of intriguing responses to these challenges. Because some of these defenses of representation can also be seen as challenging some of my own skeptical claims, it is important to examine them closely to see if they rescue the representational posits from my critique. I argue that they fail to do this, and that if anything they help show just why certain notions are ill-suited for cognitive modeling. Finally, I address some of the ramifications of the arguments presented in the earlier chapters. If many representational notions now employed in cognitive research are, as I suggest, not representational at all, then we need to rethink the extent to which these newer accounts are really so different from the "pre-cognitivist," behaviorist theories of psychological processes. I suggest that some

behaviorists, like Hull, often proposed internal mediational states that were not significantly different, in terms of functionality, from what today gets described in representational terms. A second implication of my arguments concerns the status of folk psychology. If I'm right, then many models of cognitive processes currently being proposed do not actually appeal to inner representational states. Because commonsense psychology is deeply committed to mental representations, the truth of these theories would entail eliminative materialism, the radical thesis that folk psychology is fundamentally wrong and states like beliefs and desire do not actually exist. In the final section of this chapter, I'll sketch one way this might come about that is not as preposterous as it initially sounds.

This book has taken a long time to complete and I have received a great deal of help along the way from numerous colleagues, students and friends. Among those providing helpful criticisms, insights and suggestions are William Bechtel, Tony Chemero, Marian David, Neil Delaney, Michael Devitt, Steve Downes, Chris Eliasmith, Keith Frankish, Carl Gillett, Terry Horgan, Todd Jones, Lynn Joy, Matthew Kennedy, Jaegwon Kim, John Schwenkler, Matthias Scheutz, Peter Godfrey-Smith, Stephen Stich, and Michael Strevens. I'm especially grateful to Robert Cummins, Fred Dretske, Keith Frankish, Tony Lambert, Leopold Stubenberg, Fritz Warfield, and Daniel Weiskopf who read substantial portions of earlier drafts of the manuscript and provided extremely helpful suggestions. I also want to thank Ryan Greenberg and Kate Nienaber who did the illustrations that appear in the final chapter, and my sister, Julie Talbot, who rendered some much-needed proofreading of the entire manuscript. Hilary Gaskin of Cambridge University Press provided everything an author can hope for from an editor, and Susan Beer made the copy-editing remarkably simple and straightforward. I should also acknowledge the many climbing partners who over the years, on endless drives and at cramped belay stances, humored me as I tried out some of the ideas that appear here – I imagine that occasionally one or two considered cutting the rope.

Some of the arguments presented here have appeared in a different context in other published works, most notably in "Are Receptors Representations?" (2003, *Journal of Experimental and Theoretical Artificial Intelligence* 15: 125–141); "Do Connectionist Representations Earn Their Explanatory Keep?" (1997, *Mind and Language* 12 (1): 34–66), and "Rethinking Distributed Representation" (1995, *Acta Analytica* 14: 9–25). I have also benefited a great deal from feedback from audiences at the University of Utah, the University of Cincinnati, The University of Nevada, Las Vegas, the University of Notre Dame, the Southern Society

of Philosophy and Psychology Annual Meeting (2005, Durham, NC); Cognitive Science in the New Millennium Conference (2002, Cal. State Long Beach), Society for Psychology and Philosophy Annual Meeting (1994, Memphis, Tennessee), and the IUC Conference on Connectionism and the Philosophy of Mind (1993, Bled, Slovenia). I am extremely grateful to the University of Notre Dame for awarding me with an Associative Professor's Special Leave to complete this book. I would also like to thank my department chair, Paul Weithman, who has been especially supportive of this project in a variety of different ways.

Finally, I would like to offer a special thanks to Stephen Stich, whose support and advice over the years has always proven invaluable. Nearly twenty-five years ago, he presented a devastating challenge to the received view that cognitive processes require mental representations (Stich 1983). Since no other person has had as much of an impact on my philosophical career, it is perhaps not surprising that, despite significant changes in cognitive research and the philosophy of mind, I find myself a quarter century later promoting views that are in much the same skeptical spirit.