1

Demands on a representational theory

A common feature of scientific revolutions is the discarding of the theoretical posits of the older theory in favor of the posits invoked by the new theory. The abrupt shift in the theoretical ontology is, of course, one of the things that can make a scientific upheaval so dramatic. Sometimes, however, it happens that the displaced posits hang around for a considerable stretch of time. Despite losing their explanatory value, they nevertheless retain their stature and prominence as even revolutionary thinkers resist abandoning something central to their basic understanding of the subject. The posit is perhaps transformed and re-worked as theorists contrive to fit it into a new explanatory framework for which it is ill-suited. Yet its appearance in the new theory is motivated not by any sort of explanatory necessity, but by a reluctance to reject familiar ontological commitments. When this happens, there can be a number of undesirable consequences. One is a failure to appreciate just how radical the new theoretical framework is; another is a confused understanding of the explanatory framework of the new theory, due to an extended attempt to incorporate theoretical posits that don't belong.

The status of celestial spheres shortly after the Copernican revolution helps illustrate this point. In Ptolemy's system, the spheres did real explanatory work; for instance, they helped explain what kept the massive array of stars in place as they orbited around the Earth. Without some sort of "starry vault" to anchor the stars as they rotated, they would inevitably lose their relative positions and we would look up to a different sky every night. The solid spheres provided the secure medium to prevent this from happening. But with the new Copernican cosmology, the stars stopped moving. Instead, it was the Earth that rotated, spinning on a 24-hour cycle and creating the false impression of revolving stars. Consequently, a central assumption that supported the need for celestial spheres was dropped from the new model, and it became possible to view the stars as stationary points in empty space. And yet, Copernicus and others refused to abandon the idea of semi-solid spheres housing not only the stars, but the different planets as well. This reluctance to discard the spheres from the new cosmology was no doubt due to considerations that went substantially beyond science. Historical, theological, cultural, and perhaps even "folk" considerations all played an important role in preserving the spheres, despite increasing problems in making them conform to the new theory. Although Tycho Brahe recommended abandoning solid spheres, Kepler rescued them as semi-abstract posits that he felt were essential for understanding the celestial system. It wasn't until Descartes's re-conceived space as a giant container that people let go of the idea of a starry vault (Crowe 2001; Donahue 1981).

The central theme of this book is that something very similar is currently taking place in our scientific understanding of the mind. In cognitive science, there has been something like a central paradigm that has dominated work in psychology, linguistics, cognitive ethology and philosophy of mind. That paradigm is commonly known as the classical computational theory of cognition, or the CCTC for short.[1] At the heart of the classical paradigm is its central explanatory posit – internal symbolic representations. In fact, the notion of internal representation is the most basic and prevalent explanatory posit in the multiple disciplines of cognitive science. The representational underpinning of cognitive science is, as one author puts, "what the theory of evolution is to all of biology, the cell doctrine to cellular biology, the notion of germs to the scientific concept of disease, the notion of tectonic plates to structural geology" (Newell 1980, p. 136). In the minds of many psychologists, linguists, ethologists and philosophers, the positing of internal representations is what *makes* a given theory cognitive in nature.

However, in the last two decades there have been several radical theoretical departures from the classical computational account. Connectionist modeling, cognitive neuroscience, embodied cognitive accounts, and a host of other theories have been presented that offer a very different picture of the architecture and mechanisms of the mind. With new processes like "spreading activation," "distributed constraint satisfaction," and "stochastic-dynamical processes," the operations of what John Haugeland (1997) has referred to as "new fangled" AI systems don't have much in common

[1] It is also sometimes called "GOFAI" for "Good-Old-Fashioned-Artificial-Intelligence," the "Physical Symbol Hypothesis," the "Computer Model of the Mind" (CMM), "Orthodox Computationalism," the "Digital Computational Theory of Mind" (DCTM), and a host of other names. There are now so many labels and acronyms designating this class of theories that it is impossible to choose one as "the" accepted name.

with the familiar symbol-based approach of the classical paradigm. Yet despite massive differences between classical accounts and the newer theories, the latter continue to invoke inner representations as an indispensable theoretical entity. To be sure, the elements of the newer theories that are characterized as representations look and act very differently than the symbols in the CCTC. Nevertheless, the new accounts share with conventional computational theories the basic idea that inner structures in some way serve to stand for, designate, or mean something else. The commitment to inner representations has remained, despite the rejection of the symbol-based habitat in which the notion of representation originally flourished.

My aim is to argue that this is, for the most part, a mistake. A central question I'm going to address in the following pages is, "Does the notion of inner representation do important explanatory work in a given account of cognition?" The answer I'm going to offer is, by and large, "yes" for the classical approach, and "no" for the newer accounts. I'm going to suggest that like the notion of a starry vault, the notion of representation has been transplanted from a paradigm where it had real explanatory value, into theories of the mind where it doesn't really belong. Consequently, we have accounts that are characterized as "representational," but where the structures and states called representations are actually doing something else. This has led to some important misconceptions about the status of representationalism, the nature of cognitive science and the direction in which it is headed. It is the goal of this book to correct some of these misconceptions.

To help illustrate the need for a critical analysis like the one I am offering, try to imagine what a non-representational account of some cognitive capacity or process might look like. Such a thing should be possible, even if you regard a non-representational account as implausible. Presumably, at the very least, it would need to propose some sort of internal processing architecture that gives rise to the capacity in question. The account would perhaps invoke purely mechanical operations that, like most mechanical processes, require internal states or devices that in their proper functioning go into particular states when the system is presented with specific sorts of input. But now notice that in the current climate, such an account would turn out to be a representational theory after all. If it proposes particular internal states that are responses to particular inputs, then, given one popular conception of representation, these would qualify as representing those inputs. And, according to many, any functional architecture that is causally responsible for the system's performance can be characterized as encoding the system's knowledge-base, as implicitly representing the system's know-how. If we accept current attitudes about the nature of cognitive representation, a non-representational, purely mechanistic account of our mental capacities is not simply implausible – it is virtually *inconceivable*. I take this to be a clear indicator that something has gone terribly wrong. The so called "representational theory of mind" should be an interesting empirical claim that may or may not prove correct; representations should be unique structures that play a very special sort of role. In many places today, the term "representation" is increasingly used to mean little more than "inner" or "causally relevant" state.

Returning for a moment to our analogy between celestial spheres and representation, it should be noted that the analogy is imperfect in a couple of important ways. First, in the case of the spheres, astronomers had a fairly good grasp of why they were needed in Ptolemy's system. By contrast, there has been much less clarity or agreement about the sort of role the notion of representation plays in cognitive science theories in general, including the older paradigm. Thus, one of my chores will be to sort out just how and why such a notion is needed in the CCTC. A second dis-analogy is that in the case of the spheres, there was, for the most part, a single notion at work and it was arguably that same notion that found its way into Copernicus's system. However, in the case of representation, there are actually a cluster of very distinct notions that appear in very distinct theories. Most of these notions are based on ideas that have been around for a long time and certainly pre-date cognitive science. Some of these notions, when embedded in the right sort of account of mental processes, can play a vital role in the theory. Other notions are far more dubious, at least as explanatory posits of how the mind works. My claim will be that, for the most part, the notions that are legitimate – that is, that do valuable explanatory work – are the ones that are found in the CCTC. The notions of representation that are more questionable have, by and large, taken root in the newer theories. I propose to uproot them.

Methodological matters

The goals of this book are in many ways different from those of many philosophers investigating mental representation. For some time philosophers have attempted to develop a naturalistic account of intentional content for our commonsense notions of mental representation – especially our notion of belief. By "naturalistic account" I mean an account that explains the meaningfulness of beliefs in the terms of the natural sciences, like physics or biology. The goal has been to show how the representational character of our beliefs can be explicated as part of the natural world. While many of these accounts are certainly inspired by the different ways researchers appeal to representation in cognitive theories, they neither depend upon nor aim to enhance this research. Instead, the work has been predominantly conceptual in nature, and the relevant problems have been of primary interest solely to philosophers.

By contrast, my enterprise should be seen as one based in the philosophy of science – in particular, the philosophy of cognitive science. The goal will be to explore and evaluate some of the notions of representation that are used in a range of cognitive scientific theories and disciplines. Hence, the project is similar to that, say, of a philosopher of physics who is investigating the theoretical role of atoms, or a philosopher of biology exploring and explicating competing conceptions of genes. This way of investigating mental representation has been explicitly adopted and endorsed by Robert Cummins (1989) and Stephen Stich (1992). Cummins's explanation of this approach is worth quoting at length:

> It is commonplace for philosophers to address the question of mental representation in abstraction from any particular scientific theory or theoretical framework. I regard this as a mistake. Mental representation is a theoretical assumption, not a commonplace of ordinary discourse. To suppose that "commonsense psychology" ("folk psychology"), orthodox computationalism, connectionism, neuroscience, and so on all make use of the same notion of representation is naive. Moreover, to understand the notion of mental representation that grounds some particular theoretical framework, one must understand the explanatory role that framework assigns to mental representation. It is precisely because mental representation has different explanatory roles in "folk psychology," orthodox computationalism, connectionism, and neuroscience that it is naive to suppose that each makes use of the same notion of mental representation. We must not, then, ask simply (and naively) "What is the nature of mental representation?"; this is a hopelessly unconstrained question. Instead, we must pick a theoretical framework and ask what explanatory role mental representation plays in that framework and what the representation relation must be if that explanatory role is to be well grounded. Our question should be "What must we suppose about the nature of mental representation if orthodox computational theories (or connectionist theories, or whatever) of cognition are to turn out to be true and explanatory?" (1989, p. 13)

Cummins's own analysis of representation in classical computational theory will be discussed in some detail in chapter 3, where I will offer modifications to his account. For now, I want to appeal to the Cummins model to make clear how my own account should be understood. My analysis is very much in the same spirit as what Cummins suggests, but with a couple of caveats. First, Cummins and Stich seem to assume that to demarcate the different notions of representation one should focus upon the theory in which the notion is embedded. However, a careful survey of cognitive research reveals that the same core representational notions appear in different theories and different disciplines. Hence, a better taxonomy would be one that cuts across different theories or levels of analysis and classifies types of representational notions in terms of their distinctive characteristics. Toward the end of this chapter, I'll explain in more detail the demarcation strategy I plan to use. Second, Cummins doesn't mention the possibility that our deeper analysis might discover that the notion of representation invoked in a theory actually turns out to play *no* explanatory role. Yet I'll be arguing that this is precisely what we do find when we investigate some of the more popular accounts of cognition commonly characterized as representational in nature.

Because the expanse of cognitive science is so broad, my analysis cannot be all-encompassing and will need to be restricted in various ways. For instance, my primary focus will be with theories that attempt to explain cognition as something else, like computational or neurological processes. In such theories, researchers propose some sort of process or architecture – a classical computational system or a connectionist network – and then attempt to explain cognition by appealing to this type of system. In these accounts, talk of representation arises when structures inherent to the specific explanatory framework, like data structures or inner nodes, are characterized as playing a representational role. Theories of this sort are reductive in nature because they not only appeal to representations, but they identify representations with these other states or structures found in the proposed framework. This is to be contrasted with psychological theories that appeal to ordinary notions of mental representation without pretending to elaborate on what such representation might be. For example, various theories simply presuppose the existence of beliefs and concepts to account for different dimensions of the mind, offering no real attempt to further explain the nature of such states, or representation in general. I'll be more concerned with theories that invoke representations as part of an explanatory system and at the same time offer some sense of what internal representations actually are.

Since my aim is to assess critically the notion of representation in cognitive theories, I won't be arguing for or against these theories themselves, apart from my evaluation of how they use a notion of representation. The truth or falsehood of any of these theories is, of course, an empirical matter that will depend mostly on future research. Even when I claim that a cognitive theory employs a notion of representation that is somehow bogus, or is treating structures as representations that really aren't, I don't intend this to suggest that the theory itself is utterly false. Instead, I intend it to suggest that the theory needs conceptual re-working because it is mis-describing a critical element of the system it is trying to explain.

Still, even this sort of criticism raises an important question about the role of philosophy in empirical theory construction. Why should a serious cognitive scientist who develops an empirical theory of cognition that employs a notion of representation pay attention to an outsider claiming that there is something wrong with the notion of representation invoked? What business does a philosopher have in telling any researcher how to understand his or her own theory? My answer is that in the cross-disciplinary enterprise of cognitive science, what philosophers bring to the table is a historical understanding of the key notions like representation, along with the analytic tools to point out the relevant distinctions, clarifications, implications, and contradictions that are necessary to evaluate the way this notion is used (and ought not to be used). To some degree, our current understanding of representation in cognitive science is in a state of disarray, without any consensus on the different ways the notion is employed, on what distinguishes a representational theory from a non-representational one, or even on what something is supposed to be doing when it functions as a representation. As psychologist Stephen Palmer notes, "we, as cognitive psychologists, do not really understand our concepts of representation. We propose them, and talk about them, argue about them, and try to obtain evidence in support of them, but we do not understand them in any fundamental sense" (Palmer 1978, p. 259). It is this understanding of representation, in a fundamental sense, that philosophers should help provide.

One reason for the current state of disorder regarding representation is that it is a theoretical posit employed in an unusually broad range of disciplines, including the cognitive neurosciences, cognitive psychology, classical artificial intelligence, connectionist modeling, cognitive ethology, and the philosophy of mind and psychology. This diversity multiplies when we consider the number of different theories within each of these disciplines that rely on notions of representation in different ways. It would be impossible to examine all of these different theoretical frameworks and applications of representational concepts. Hence, the overall picture I want to present will need to be painted, in spots, with broad strokes and I'll need to make fairly wide generalizations about theories and representational notions that no doubt admit of exceptions here and there. This is simply an unavoidable part of doing this type of philosophy of science, given the goal

Cambridge University Press
978-0-521-85987-5 - Representation Reconsidered
William M. Ramsey
Excerpt
More information

of providing general conclusions about a diverse array of trends and theories on this topic. If what I say does not accurately describe your own favorite theory or model, I ask that you consider my claims in light of what you know about more general conventions, attitudes, assumptions and traditions.

If I am going to establish that certain notions of representation in cognitive science are explanatorily legitimate while others are not, we need to try to get a better sense of what constitutes "explanatory legitimacy." Given the current lack of agreement about representation, figuring out just how such a notion is supposed to work in a theory of mental processes is far from easy. Despite the large amount of material written on mental representation over the years, it is still unclear how we are supposed to think about it. As John Searle once noted, "There is probably no more abused a term in the history of philosophy than 'representation' . . ." (1983, p. 11). Arguably, the same could be said about "representation" in the history of cognitive science. What does the positing of internal representations amount to? When is it useful to do so and when is it not? Exactly what is being claimed about the mind/brain when it is claimed to have representational states? Answering these questions is, in large measure, what this book will try to do. As a first pass, it will help to first step back and consider in more general terms some of our ordinary assumptions and attitudes about representational states.

1.1 REPRESENTATION AS CLUSTER CONCEPT(S)

Cognitive researchers often characterize states and structures as representations without a detailed explication of what this means. I suspect the reason they do this is because they assume they are tapping into a more general, pre-theoretical understanding of representation that needs no further explanation. But it is actually far from clear what that ordinary conception of representation involves, beyond the obvious, "something that represents." Perhaps the first thing we need to recognize is that, as others have pointed out (Cummins 1989; von Eckardt 1993), it is a mistake to search for *the* notion of representation. Wittgenstein famously suggested that concepts have a "family-resemblance" structure, and to demonstrate his point, he invoked the notoriously disjunctive notion of a game. But Wittgenstein could have just as easily appealed to our ordinary notion of representation to illustrate what he had in mind. We use the term "representation" to characterize radically different things with radically different properties in radically different contexts. It seems plausible that our notion of representation is what is sometimes called a "cluster" concept (Rosch and Mervis 1975; Smith and Medin 1981) with a constellation of different types that share various nominal features, but with no real defining essence. If this is the case, then one popular philosophical strategy for exploring representation in cognitive science is simply untenable.

When trying to understand representation in cognitive science, writers often offer semi-formal, all-encompassing definitions that are then used as criteria for determining whether or not a theory invoking representations is justified in doing so. Initially, this might seem like a perfectly reasonable way to proceed. We can simply compare the nature of the posit against our crisp definition and, with a little luck, immediately see whether the alleged representation makes the cut. However, I believe this strategy has a number of severe flaws. First, in many cases the definition adds more mystery and confusion that it clears away. For example, Newell has famously defined representation in terms of a state's capacity to designate something else, and then defines designation in this way: "An entity X designates an entity Y relative to a process P, if, when P takes X as input, its behavior depends on Y" (1980, p. 156).

It is far from clear how this definition is supposed to refine our understanding of designation or representation. After all, my digestive processes sometimes takes a cold beer as input and when it does so its behavior often depends on whether or not I've had anything else to eat, along with a variety of other factors. Does this mean a cold beer designates my prior food intake? Presumably not, yet it appears the definition would say that it does. Newell clearly intends to capture a relation between X, P and Y that is different from this, yet the definition fails to explicate what this relation might be.

Second, virtually all of the definitions that have been offered give rise to a number of intuitive counter-examples. As we have just seen, Newell's criteria, taken as sufficient conditions, would suggest that a beer I've ingested serves a representational function, which it clearly does not. As we will see in the forthcoming chapters, similar problems plague the definitions offered by other writers who propose definitions of representation. Counter-examples come in two forms – cases that show a proposed definition is too inclusive (i.e., treat non-X as if they are Xs) and cases that show a proposed definition is too exclusive (i.e., treat actual Xs as if they are not Xs). Definitions of representation typically fail because of the former sort of counter-examples – states and structures that play no representational role are treated as if they actually do.

Now it might be thought that these difficulties are simply due to a bunch of flawed definitions, while the original goal of constructing a general definition for representation is still worth pursing. Yet the research on categorization judgments suggests there is reason to think these problems run deeper and are symptomatic not of bad analysis, but of the nature of our underlying pre-theoretical understanding of representation. If Rosch and various other psychologists are correct about the disjunctive way we encode abstract concepts, then the difficulties we see with these definitions are exactly what we should expect to find. Simple, tidy, conjunctive definitions will always fall short of providing a fully satisfactory or intuitive analysis. They might capture one or two aspects of some dimension of our general understanding, but they won't reveal the multi-faceted nature of how we really think about representation.

Suppose these psychologists are right about our conceptual machinery and that our concept of representation is itself a representation of an array of features clustered around some sort of prototype or group of prototypes. This would make any crisp and tidy definition artificial, intuitively unsatisfying, and no better than a variety of other definitions that would generate very different results about representation in theories of the mind. If we want to evaluate the different notions of representation posited in scientific theories, a more promising tack would be to carefully examine the different notions of representation that appear in cognitive theories, get as clear as possible about just what serving as a representation in this way amounts to, and then simply ask ourselves – is this thing really functioning in a way that is recognizably representational in nature? In other words, instead of trying to compare representational posits against some sort of contrived definition, we can instead compare it directly to whatever complex concept(s) we possess to see what sort of categorization judgment is produced. If, upon seeing how the posit in question actually functions we are naturally inclined to characterize its role as representational in nature, then the posit would provide us with one way of understanding how physical systems can have representations. If, on the other hand, something is functioning in a manner that isn't much like what we would consider to be a representational role, then the representational status of the posit, along with its embedding theory, is in trouble. This is roughly how my analysis will proceed – by exploring how a representational posit is thought to operate in a system, and then assessing this role in terms of our ordinary, intuitive understanding of what a representation is and does. To some degree, this means our analysis will depend on a judgment call. If this is less tidy than we would like, so be it. I would prefer a messier analysis that presents a richer and more accurate account of representation than one that is cleaner but also off the mark. Eventually, we may be able to