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 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 construct something like a general analysis or theory of representation. But this can only happen after first exploring the ways in which physical structures may or may not accord with our more basic, intuitive understanding of representation.

A very different question worth considering is this: why should we care if a given representational posit accords with our commonsense understanding of representation in the first place? If these are technical, scientific posits, what difference does it make whether the theorist uses the term "representation" to refer to things that behave in a manner sanctioned by intuition? Isn't is really just the explanatory value of a theoretical posit that matters? And if so, isn't it trivially true that cognitive systems use representations? An illustration of this attitude is provided by Roitblat (1982), who happily proclaims that, "[t]o assume the existence of a representation is rather innocuous and should rarely be an issue for theoretical dispute" (Roitblat 1982, p. 355). Since Roitblat defines representation as *any* internal change caused by experience, it is not surprising that he thinks it pointless to wonder about their existence.

However, I actually think that quite a lot rides on whether or not a representational posit actually functions in a way that we are naturally inclined to recognize as representational. First, it is important to think carefully about what it means to say that a given notion is doing important explanatory work. Suppose someone claims to have a representational theory of diseases, and posits representational states as the cause of most illnesses. Upon further analysis, we discover that the theorist is simply using the term "representation" to refer to ordinary infectious agents, like viruses and bacteria. Moreover, we discover that there is nothing intuitively representational about the role the theory assigns to these agents – they just do the things infectious agents are ordinarily assumed to do. Notice how silly it would be for the theorist to defend his representational account by pointing out that he isn't interested in our ordinary notion of representation, and that what matters is that his representational posits do important explanatory work. While the posits would indeed do explanatory work, they wouldn't actually be serving as *representational* posits. This would not be a case where a technical notion of representation is playing some explanatory role. Instead, this would be a scenario where a notion of representation would not be playing *any* explanatory role; it would be completely absent from the theory. All of the work would be done by ordinary notions of infectious agents. This is because there is nothing about the job these states are doing that is intuitively recognizable as representational in nature. Unless a posit is in *some* way grounded in our ordinary understanding of representation, it is simply not a representational posit, in any sense.

In earlier work (Ramsey 1997), I chose not to address the issue of whether or not a proposed form of representation actually was a representation, and instead focused on the question of explanatory utility, asking if a notion of representation did any explanatory work. I now think this was a mistake. It was a mistake because what matters is not explanatory work, but explanatory work *qua* representation. In showing that a posit fails to do explanatory work *qua* representation, what is typically shown is that the proposed posit doesn't function in a representational manner; that is, it is not a representation after all. So the metaphysical issues cannot be avoided, even if one's primary interest is with questions of explanatory utility. There are different ways a theoretical posit from an older framework can be mistakenly retained in a new framework. One way, suggested by the case of the crystal spheres, is if the posit fails to correspond to anything in the new ontology. But another way is if some part of the new ontology is characterized as playing the role associated with the old posit when in truth, it is playing a completely different role. That is what I will claim is happening with the notion of representation.

Second, the positing of inner representations typically comes with a lot of assumptions, expectations, concerns, inferential entitlements, and other theoretical attachments that are rooted in (and licensed by) our ordinary ways of thinking about representation. The significance of a representational theory of the mind stems in large measure from the different elements that are associated with representational states as ordinarily understood. For example, when theorists posit inner representations, they typically assume that they now have an important way to explain how the system can fail to behave appropriately. It is now possible to explain faulty behavior as sometimes stemming from false representations of the world. In fact, considerable philosophical effort has been devoted to explaining how it is actually possible for a physical state to be in error – to misrepresent the nature of reality. This is an important topic because the possibility of misrepresentation is built into our ordinary way of understanding what it is to represent. If someone announced that they were using a technical notion of representation that didn't admit of misrepresentation, we would not think that this is just another way of handling the problem of error. Instead, we would think that whatever the posited state was doing, it wasn't playing a representational role. We can't posit representational states to do many of the things they are supposed to do in a theory unless the posit itself is sufficiently similar to the sort of things we pre-theoretically think representations are.

This last point also helps us see that there is more at stake here than a mere terminological or semantic squabble. With a simple terminological mistake, a non-A is mistakenly called an "A," though it is not ascribed any of the features normally associated A. This might happen when someone is learning a language. In the case of real conceptual confusion, on the other hand, a non-A is called an "A" and also treated as having all (or most) of the features normally associated with A. It is clearly one thing to mistakenly think the word "dog" refers to cats, it is quite another thing to mistakenly think that dogs are a type of cat. The confusion I will be addressing involves the latter sort of mistake – people thinking that non-representational states and structures really are a type of representation. This leads them to make the further mistake of thinking that the sort of conceptual linkages and accompaniments associated with representation should be ascribed to non-representational entities.

Finally, contrary to Roitblat's claim, the question of whether or not the brain performs cognitive tasks by using inner representations is an important one that deserves to be investigated with the same seriousness that we investigate other important empirical questions. Notice how many traditional problems could be resolved by just ignoring our intuitive understanding of things and instead offering new definitions. Can machines be conscious? Well, let's just define consciousness as electrical activity and thereby prove that they can. Do non-human primates communicate with a language? Sure, if we think of language as any form of communication. Does smoking really cause lung cancer? No, not if we ignore our ordinary way of thinking about causation and employ a technical notion where to be a cause is to be a necessary and sufficient condition. Most of us would treat this sort of strategy for addressing these questions as uninteresting ploys that dodge the real issues. Similarly, any suggestion that we should answer the question, "does this system employ inner representations?" in a manner that ignores our intuitive understanding of what a representation is and does is equally misguided. Of course, this doesn't mean that there can't be notions of representation that are somewhat technical, or that depart to *some* degree from our folk notions of mental representation. In fact, as we will see in chapter 3, notions of representation used in classical computational accounts of cognition are both valuable and somewhat unique to that explanatory framework. What it does mean, however, is that the theoretical notions of representation must overlap sufficiently with our pre-theoretical understanding so that they function in a way that is, indeed, recognizably representational in nature.

When I suggested earlier that our ordinary conception of representation cannot be captured by simple definitions, I did not mean to imply that it can't be illuminated in various ways. If our notion of representation involves a cluster of features, we can ask what some of those features are. In fact, a strong case can be made that there is not one cluster but two overlapping constellations, corresponding with two different families of representational notions. One cluster corresponds to various notions of mental representation, the other to different types of non-mental representation. Cognitive scientists and philosophers often tap into these clusters when they construct theories about the mind that appeal to representations, and as we will see throughout our discussion, the non-mental cluster is often used to explicate cognitive representation. Consequently, it will help to briefly look at some of the aspects of these families of representational notions to get a better sense of where the more scientific notions of cognitive representation come from.

1.1.1 Mental representation within folk psychology

Our ordinary, "folk" conception of mental representation includes things like different types of knowledge, propositional attitudes (beliefs, desires, hopes, etc.), memories, perceptual experiences, ideas, different sorts of sensations, dream states, imaginings, and various emotional responses to circumstances. Some of these notions are clearly closer to what might be considered the "center" of the cluster than others. In particular, our notions of basic *thoughts* – propositional attitudes[2] – appear to be more central to our ordinary understanding of mental representation and most writers treat them as paradigmatic. I'll focus on thoughts in my discussion here (or more accurately, on our conception of thoughts), though a great deal of what I'll say generalizes to other notions of mental representation as well. So, what do we take to be the basic features of thoughts?

It might be supposed that explaining our commonsense perspective on thoughts and other mental representations should be a trivial and uncontroversial affair. Ex hypothesis, our ordinary attitudes about mentality are common knowledge and its main features are easily accessible to all. Alas, things aren't so simple. Exactly what our commonsense understanding of the mind involves and how it works is something heavily debated by both

[2] For those unfamiliar with the term, propositional attitudes are mental states such as beliefs, desires, hopes, fears, assumptions, and the like. They are, as the name implies, a certain attitude (believing, desiring, hoping, etc.) toward a proposition. Propositions are perhaps best conceived of as states of affairs.

philosophers and psychologists; at the present, there doesn't appear to be anything close to an emerging consensus. Since these different accounts of commonsense psychology entail different accounts of how we regard mental representations, it is difficult to articulate this commonsense notion without stepping on someone's toes.

On one side of this debate are many philosophers and psychologists, including myself, who maintain that our commonsense or folk psychology functions as a predictive and explanatory *theory* (Churchland 1981, 1989; Gopnik and Wellman 1992; Stich and Nichols 1993). This view – the "theory-theory" – suggests that, like any theory, commonsense psychology is comprised of both theoretical posits and a number of law-like generalizations. The main posits include various representational states like beliefs, desires and other propositional attitudes, as well as various qualitative states like pains. The "laws" of folk psychology are the platitudes we use to predict and explain one another's behavior. Thus, on most versions of the theory-theory, we treat mental states like beliefs as entering into causal relations that support a wide range of generalizations. One of the more controversial aspects of the theory-theory is that it opens up the possibility of eliminativism – the view that folk psychology might be a radically false theory, and that we will come to discover that its posits, like beliefs and desires, don't actually exist.

However, not everyone accepts the theory-theory account of our ordinary understanding of the mind. Some reject it because they regard belief-desire psychology to be something very different from a system that posits inner causes and law-like generalizations. On one view, it is a way of making sense of the activities of rational and linguistic agents, used to classify and identify rather than to explain and predict. As one author puts it, "[F]olk psychology, so called, is not a body of a theory but an inherited framework of person-involving concepts and generalizations" (Haldane 1993, pp. 272–273). Others reject the theory-theory by claiming that to explain and predict behavior, we rely not on a theory but on a type of simulation. According to this view, we take some of our own information-processing mechanisms "off-line" (so the mechanism generates predictions instead of behavior) and then feed it relevant pretend beliefs and desires that are assumed to be held by the agent in question. Then, sub-consciously, we use our own decision-making mechanisms to generate output which can thereby serve as predictions (and, in other circumstances, explanations) of the agent's behavior. No theoretical posits or laws – just the use of our own machinery to grind out recommended actions that we can then exploit in accounting for the behavior of others (Gordon 1986; Goldman 1992).

Hence, there is considerable disagreement about what our commonsense psychology is really like, which in turn leads to disagreement about what our concepts of mental representation are like. Indeed, there is even disagreement about how we ought to *figure out* what commonsense psychology is really like (Ramsey 1996). So much for the commonality of commonsense! Of course, in presenting our conception of mental representations, there is no way that we can hope to resolve all of these debates here. But for now, given that we just want to get the ball rolling, perhaps we don't need to resolve all of them. Despite the different disputes about the nature of commonsense psychology, there is little disagreement over whether we actually *have* commonsense notions of mental representation. So perhaps there are some basic features associated with those notions that can be agreed upon by most. I think there are at least two.

Intentionality

Most philosophers agree that our concepts of mental representations involve, in some way, intentionality (also referred to as the "meaning," "intentional content," or the "semantic nature" of mental representations). Intentionality (in this context) refers to "*aboutness*."[3] Thoughts, desires, ideas, experiences, etc. all *point to* other things, though they could also, it seems, point to themselves. Intentionality is this feature of pointing, or designating, or being about something. Typically, mental representations are about a variety of types of things, including properties, abstract entities, individuals, relations and states of affairs. My belief that Columbus is the capital of Ohio is about Ohio, its seat of government, the city of Columbus, and the relation between these things. On most accounts, we treat the intentional nature of our thoughts as crucial for their individuation; that is, we distinguish different thoughts at least in part by appealing to what they are about. My belief about the capital of Ohio is clearly a different mental state than my belief about the capital of Indiana. In this way, intentionality serves as a central, distinguishing feature of all mental representations. It is hard to see how something could qualify as a mental representation in the ordinary sense unless it was *about* something – unless it in some way stood for something else.

On most accounts, the intentionality of mental representations is an extremely unique feature of minds and minds alone. While public signs and linguistic symbols are meaningful, their meaning is generally assumed to be derivative, stemming from the conventions and interpretations of

[3] This helpful way of characterizing intentionality is from Dennett and Haugeland (1987).

thinking creatures. That is, the aboutness of a word or road sign is thought to exist only through the aboutness of our thoughts – in particular, the aboutness of the thought that these physical shapes stand for something else. Only thoughts and other mental representations are assumed to have what is called "original" or "intrinsic" intentionality. Intuitively, no one needs to *assign* a meaning to my thought that Columbus is the capital of Ohio for it to be the case that the capital of Ohio is what that thought is about. Such a thought seems to be, as one philosopher has put it, a sort of "unmeant meaner"[4] – a state whose meaning is not derived from other sources. How this is possible is often assumed to be one of the great mysteries associated with mentality.

Along with this "intrinsicality," the intentionality we associate with mental representations brings with it a number of other curious features that have received considerable attention, especially from philosophers of mind. For example, the intentional relation between representation and what it represents is odd in that the latter need not actually exist. For most sorts of relations, both relata are needed for the actual relation to obtain. Yet we can have thoughts about non-existent entities like unicorns and Sherlock Holmes, suggesting the nature of the intentional relation between thoughts and their objects is highly atypical. Furthermore, thoughts can represent the world as being a way that it isn't. Beliefs can be false, perceptual illusions misrepresent reality, and our hopes and desires entertain states of affairs that may never come about. How this is possible is far from obvious. And there is also the curious feature of intentionality referred to as "opacity." Although thoughts are individuated in terms of what they are about, two thoughts about the same state of affairs are not treated as identical. Even though I can be characterized as believing that John Wayne was an actor, I can't be said to believe that Marion Morrison was an actor even though, as it turns out, John Wayne was actually Marion Morrison. The different *ways* we can represent things and events matters a great deal for our ordinary understanding of mental representation.

The oddness of the intentionality we associate with thoughts has led some, most famously Brentano, to suggest that the mind is in some way non-physical. The intentional nature of mental representations is sometimes characterized as an "irreducible" property – a feature that cannot be explained through the natural sciences. Since most contemporary philosophers of mind are physicalists, a major project over the last thirty years has

[4] Dennett 1990. It should be noted that Dennett rejects the idea of intrinsic intentionality and employs the phrase "unmeant meaner" in jest.

been to try to show how we *can*, in fact, explain intentionality in physical terms. Many of these attempts to appeal to the sort of features associated with non-mental representations we will look at in section 1.1.2. While there is considerable debate about how best to explain intentionality, there is near unanimity on the central role it plays in our commonsense understanding of mental representations. Indeed, its importance is so central that there seems to be a tacit assumption held by many philosophers that a theory of intentionality *just is* a theory of representation. As we will see below, this assumption is, for a variety of reasons, highly questionable.

Causality

The second sort of relatively uncontroversial feature associated with mental representations is the set of causal relations that commonsense assigns to our thoughts. Intuitively, mental representations are states that *do* various things. Although philosophers once denied that thoughts could serve as causes (Anscombe 1957; Melden 1961), today there is general agreement that in *some* sense, our ordinary understanding of thoughts attributes to them various causal roles. For example, folk psychology treats my belief that the stove is on as a state with the content *"the stove is on"* employed in a specific range of causal relations. These relations might include being triggered by perceptual stimuli of the dial set to the "on" position, the generation of a fear that my gas bill will be too high, the production of a hand motion that turns the stove off, and so on. On one version of the theory-theory, the set of causal relations associated with our thoughts correspond to the law-like generalizations of our folk psychological theory. A popular example of such a law goes as follows: If someone wants X and holds the belief that the best way to get X is by doing Y, then barring other conflicting wants, that person will do Y. When we explain or predict behavior, the theory-theory claims we (tacitly) replace variables X and Y with whatever propositions we think an individual actually desires and believes. For instance, I might explain Joe's obsequiousness by suggesting that Joe wants a raise and believes the best way to get a raise is by complimenting the boss. This want and belief are together thought to literally cause Joe to act in the way he does. The same applies to other notions of mental representation, including desires, hopes, memories, images, and so on.

While the basic idea that mental representations partake in different causal relations is fairly straightforward and perhaps amenable to scientific treatment, there is, many would argue, a second aspect of our ordinary conception that makes the causal nature of representations more difficult to explain. It has been argued that commonsense psychology suggests that our thoughts not only interact in various ways, but that they participate in these causal relations by virtue of their content (Dretske 1988; Horgan 1989). The type of behavior a belief or desire generates is intuitively determined by what that belief or desire is about. But this makes the causal nature of mental representations more problematic. First, if the causal properties of representations depend upon their intentional properties, then all of the apparent mysteriousness of intentionality extends to their causal nature. Second, many naturalistic accounts of intentional content appeal to head-world relations and historical factors that would seem to have no way of influencing the causal powers of inner cognitive states that might be treated as representations. This has led many to abandon the idea that representations do what they do *because of* what they are about, and instead adopt the weaker position that the causal role of representations *corresponds* to their content in such a way that their interactions "make sense." If I believe that if P then Q and also come to believe P, then these two beliefs will cause me to believe Q, though not, strictly speaking, by virtue of their content (which is causally inert). In the next chapter, we will look at how this story is presented in the framework of the CCTC, while in chapter 4 we'll examine a theory that attempts to show how content *is* causally relevant.

Beyond these mundane observations about the intentionality and causality of mental representations, what little consensus there is about our commonsense picture of mentality begins to evaporate. For example, the relevance of other factors for our basic conception of mental representation, such as the role of consciousness, public language, or rationality, is far more controversial.[5] Still, it might be thought that from this very modest analysis, we have enough to begin to see what a psychological theory that appeals to inner representations ought to look like. Mental representations are states that have some sort of non-derived intentionality and that interact with other cognitive states in specific sorts of ways. Since folk psychology is arguably a primary source from which a representational perspective is derived, one could say its posits should be all that a scientific theory needs to invoke. A psychological theory that invokes inner representations is thereby a theory that invokes beliefs, desires, ideas, and other folksy notions.

[5] In fact, things are more controversial than I've even suggested here. For example, as noted above, Daniel Dennett denies that there is such a thing as original intentionality. He also rejects the idea that we treat mental representations as causes in any straightforward, billiard-ball way (Dennett 1991).

While it is true that our commonsense notions of mental representation influence psychological theorizing a great deal, it would be a mistake to assume that cognitive scientists set out simply to mimic these notions when developing their own picture of how the mind works. As we'll see in the coming chapters, researchers develop and produce theoretical notions of representation that depart in various ways from folk notions. Even when notions like belief are incorporated into scientific accounts, it is typically stretched and modified in order to fit the explanatory needs of the theory. Moreover, our ordinary notion of mental representation leaves unexplained a great deal of what a theory-builder should explain about how something actually serves as a representation. Commonsense psychology provides us with little more than a crude outline of mental representations and leaves unanswered several important questions about how representations drive cognition. This point will be addressed in greater detail in section 1.2 below. But before we examine that topic, we should also briefly consider our ordinary notions of non-mental representation.

1.1.2 Non-mental representation

As with mental representation, the commonsense class of non-mental representations is quite large and encompasses a diverse range of states and entities. These include, but are not limited to, linguistic symbols, pictures, drawings, maps, books, religious icons, traffic signals and signs, tree rings, compass needle positions, tracks in the snow, hand signals, flashing lights, and on and on. This diversity suggests that whatever non-mental representation amounts to, there are few restrictions on the types of things that qualify. Perhaps this is unsurprising if, as suggested earlier, non-mental representations all have derived intentionality. If something's status as a representation is merely assigned by minds, and if minds can assign meaning to practically anything, then we would expect there to be a very diverse array of things that serve as non-mental representations. Moreover, if non-mental representation is entirely dependent upon mental representation, it is far from clear that there is much that the former can tell us about the latter. If non-mental representations lack the central defining features associated with cognitive representations, why should we bother thinking about non-mental representations at all?

There are a couple of answers to this question. First, some have argued that it is just wrong to suppose that only mental states possess intrinsic intentionality. They have suggested that there is a type of low-level meaning "out there" in the world, possessed by physical states without the intervention of interpreting minds. For example, some authors have claimed that a tree's rings carry information about its age all by themselves, irrespective or whether or not anyone notices this (Dretske 1988). If this is correct, then it may be possible to gain some sort of insight into cognitive representation by exploring the representational character of things that are non-mental. Second, even if all non-mental representation is in some sense derivative, we might still be able to learn important facts about the nature of representation – especially about the way cognitive scientists *think about* representation – by looking at the non-mental cases. Since we are trying to gain some insight into the sort of thing researchers have in mind when they posit representations in psychological theories, it is worth at least considering the type of representations we encounter in our everyday lives.

Historically, there have been many attempts to spell out the central features of everyday representations. One such attempt is Charles Peirce's theory of semiotics (1931–58), an extremely rich but also cryptic analysis of the general nature of representation. Despite the abstruse nature of Peirce's theory, it provides at least a basic framework from which helpful insights about the character of non-mental representation can be found.[6] Since I plan to pillage the parts of Peirce's account that I find intuitively plausible, my apologies to Peirce scholars for ignoring or mistreating various nuances of his view.

One of Peirce's main contributions on representation is an analysis of the different ways in which representations – what he calls "signs" – are linked to the things they represent. Peirce appeals to three types of content "grounding"[7] relations, corresponding with three different sorts of signs. First, there are "icons," signs that are connected to their object by virtue of some sort of structural similarity or isomorphism between the representation and its object. Pictures, maps, and diagrams are all iconic representations. A picture represents a person at least in part because the former closely resembles the latter. Second, there are "indices" – signs that designate things or conditions by virtue of some sort of causal or law-like relation between the two. An array of tree rings exemplifies the category of indices since the age of the tree reliably causes the number of rings. What many philosophers today would call "natural signs" or "indicators" qualify as Peirce's indices. Pierce's third category is what he calls "symbols."

[6] See, for example, Barbara von Eckardt's (1993) excellent synopsis of Peirce's account that refines and modifies his view, highlighting its most salient and plausible components for cognitive science.

[7] Like others, I use the phrase "content grounding relation" here to designate the natural conditions or relations that are thought to give rise to the intentional relation between a representation and its object.

Symbols are connected to their objects entirely by convention. There is no further feature of a symbolic sign that bestows their content – they mean what they do entirely by stipulation. Linguistic tokens, such as written words, are paradigm cases of Peirce's symbols.

Peirce's analysis is important for our purposes because, as it turns out, these same ideas serve as the basis for different notions of representation found in cognitive science. In fact, much of what has been written about mental representation over the last thirty years can be viewed as an elaboration on Pierce's notions of icons, indices and symbols. In chapter 3, we'll look at how something quite similar to Pierce's notion of representational icons is actually an important theoretical entity in the accounts of cognition put forth in the CCTC. As we'll see, many versions of the CCTC posit inner states that serve as representations in the same sense in which the lines and figures on a map serve to represent features of some terrain. In chapter 4, we'll look at notions of representation that are based on the same sort of law-like dependencies Pierce associated with indices that appear in many of the newer accounts of cognition. In both sorts of cases, a notion of representation is put forth in accounts of cognitive processes that is based upon principles associated with our pre-theoretical understanding of non-mental representation. Our job will be to determine whether these principles can be successfully applied to cognitive states and processes in the brain so that we wind up with an explanatorily adequate account of cognitive representation.

While philosophers and cognitive scientists have attempted to explain representation by appealing to the physical relations associated with icons and indices, Peirce himself would probably regard this whole project wrong-headed. Peirce held that representation is always a triadic relation, involving (a) the sign, (b) its object, and, (c) some cognitive state of an interpreter (the "interpretant"). On Peirce's account, the interpretant is itself a representation, leading to a regress which he cheerfully embraces. But here we can simply treat the third element as a cognitive agent. For Peirce, all three elements must be involved; if any one component is missing, there is no representation. Consequently, the Peircean picture rejects any attempt to reduce representation to a dyadic relation that excludes the interpreter. For him, there can be no meaning or representational content unless there is some thing or someone *for whom* the sign is meaningful.

This makes it sound as though Peirce was a strong advocate of the original/derived intentionality distinction that, as we suggested at the outset of this section, threatens to undercut any attempt to explain mental representation in terms of what we know about non-mental representations. But it is somewhat doubtful that Peirce had in mind the original/derived intentionality distinction, since for him even mental representations have, in some sense, derived intentionality. For Peirce, *all* forms of representation involve something like an interpreter, and it is far from clear that he distinguished mental and non-mental representations in the way many do today. What *is* significant about Peirce's triadic analysis is the idea that representations are things that are *used in a certain way.* Something qualifies as a representation by playing a certain kind of role. Similarly, Peirce treats representation as a functional kind – to be a representation is to be something that does a certain job (Delaney 1993, pp. 130–156).

Peirce seems right that this is a basic feature of our ordinary understanding of representation. As Haugeland puts it, "representing is a functional *status* or *role* of a certain sort, and to be a representation is to have that status or role" (Haugeland 1991, p. 69). When we consider non-mental representations like maps, road signs, thermometers and bits of language, it is clear that these things are employed by cognitive agents as a type of tool. These are all things that serve to inform minds in some way or other about various conditions and states of affairs, and outside of that role their status as representations disappears. The proverbial driftwood washed up on an uninhabited beach does not, intuitively, represent anything, even if it happens to spell out the word "UNINHABITED BEACH" or is arranged in a way that maps a course to a nearby lake. However, if someone were to come along and use the driftwood as a type of map, then it would indeed take on a representational role.

What all of this suggests is that if our understanding of cognitive representations is based on our understanding of non-mental representations, then we need to understand how something can play a representational role *inside* a given cognitive system. If our basic notion of non-mental representation is a functional notion, like our notion of a hammer or door stop, then any cognitive theory positing inner states derived from these notions is positing states that have a job to perform. With non-mental representation, that job appears to require a full-blown cognitive agent as an employer. Exactly what that job is supposed to entail *within* a cognitive agent – in the context of psychological processes – is far from clear. If ordinary notions of non-mental representation are to form the basis for understanding representation in cognitive theories, and if those ordinary notions always presuppose some sort of representation *user*, then we need to provide some sort of account of representation use where the user isn't a full-blown mind. In the following chapters, I'll argue that we can do this for one of Peirce's signs – namely, icons – but not the others. I'll argue that things that represent in the way icons represent can be found within a mechanical or biological system, whereas this can't be done for things that represent in the manner of Peirce's symbols or indices.

While the preceding is hardly an exhaustive analysis of the nature of non-mental representation, it has highlighted two important ideas. The first is that there are basic kinds of non-mental representation and that these are also found in theories of how the mind works. Hence, theorists appeal to certain sorts of non-mental representation – discussed by Peirce – as a guide for understanding the nature of cognitive representation. The second point is that there are some *prima facie* difficulties associated with such an appeal. Central among these is the fact that our ordinary conception of non-mental representation seems to pre-suppose that they do a certain kind of job (like informing or designating), and it is far from clear how we are supposed to characterize that job without appealing to an up and running mind. Because I think this last point is extremely important (and often underappreciated) it will help to consider it more closely.

1.2 THE JOB DESCRIPTION CHALLENGE

In the last section, we were operating on the assumption that by reflecting a bit on our ordinary notions of representation, we could gain a better understanding of what it is that scientists are referring to when they claim the brain uses such states. But one might wonder why we need to look at commonsense notions of representation at all. In most scientific theories, a theoretical notion is introduced in such a manner that includes the unique properties that give the posit its specific explanatory role. The positing of genes, for example, involves a specification of the different relations and causal roles that describes the sort of job we think genes perform (Kim 1998). By using this job description, we can then go look for the actual bio-chemical structures that fit the bill. Of course, along the way we may discover that our job description needs to be modified in some way. But we cannot make any progress in understanding how a given posit is actually realized unless we first have a fairly clear understanding of what it is the posit supposedly does. This understanding is not provided by commonsense, but by the scientific theory itself.

In the case of representation, however, things are more complicated. As we've just seen, representational notions already have a home in our non-scientific conception of the world. This non-theoretical understanding constrains the sorts of things that can qualify as representational states, even in the context of a scientific theory. As we noted in section 1.1, the scientific notions must in *some* way be rooted in our ordinary conception of representation; otherwise, there would be little point in calling a neural or computational state a representation. Thus, we briefly looked at two sets of commonsense notions of representation to try to get a clearer sense of exactly what is being invoked when a theorist posits inner representations as an element of the mind. What we would like these notions to provide is a specification of the essential or core features of representation that we can then use in our assessment of scientific theories that claim to be representational. We would like something akin to a *job description* for representational posits that is analogous to what we have for other theoretical posits, like genes or protons, so that we can then determine if a given state or structure fits the bill.

Yet as we saw in the last section, an analysis of the commonsense notions doesn't really provide us with what we are after. The problem is *not* that the commonsense notions don't involve core features or offer job descriptions for representational states. Rather, the problem stems from the sort of features and roles that are associated with these notions. In the physical or biological sciences, a job description for a posit can be provided in straightforward causal/physical (or causal/bio-chemical) terms, like the pumping of ions or the production of some enzyme. But in the case of both non-mental and mental representation, the relevant roles include things like *informing*, *denoting* or *standing for something else*. It is not at all clear how these sorts of roles are supposed to be cashed out in the naturalistic, mechanistic[8] framework of a cognitive theory. Many scientific theories of the mind attempt to explain cognition in neurological or computational terms. But our ordinary understanding of representation involves features and roles that can't be translated into such terms in any obvious way.

Consider our ordinary notions of mental representation. As we saw above, our commonsense understanding of beliefs, desires, and other folk representational states assigns to them some sort of underived or intrinsic intentionality, and this feature is thought to be central to their serving as representational states. But intentionality clearly isn't a basic causal or functional property. Consequently, when we look inside a physical

[8] By "mechanistic," I simply mean an explanation that appeals to physical or perhaps what are often called "syntactic" states and processes. A mechanistic explanation accounts for some capacity or aspect of the mind by showing how it comes about through (or is realized by) structures, states and operations that could be implemented in physical processes.

system to determine if there are mental representations, it is not at all clear what we are looking for. It isn't clear what having the property of "aboutness" is supposed to entail for a state of a physical system, or how having such a feature will influence the way a physical system operates. If researchers simply adopt, without further elaboration, our ordinary notions of mental representation as part of their naturalistic accounts of the mind, we are left with an account that can't be fully understood because we have no sense of what serving as a representation in such a system is supposed to entail.

A similar problem arises with regard to our commonsense understanding of non-mental representation. Recall that here the notion of representation is associated with a user, and if we ask about the sorts of things that use such representations, the most natural answer would be a full-blown cognitive agent. Everyday examples of non-mental representations – road signs, pieces of written text, warning signals, and so on – all involve thinking agents who use the representation to stand for something else. How, then, can we specify the functional role of representation as something employed *within* cognitive systems, when it intuitively functions as something used externally *by* cognitive systems? As Dennett points out,

> nothing is intrinsically a representation of anything; something is a representation only for or to someone; any representation or system of representation thus requires at least one user or interpreter of the representation who is external to it. Any such interpreter must have a variety of psychological or intentional traits . . .: it must be capable of a variety of comprehension, and must have beliefs and goals (so it can use the representation to inform itself and thus assist it in achieving its goals). Such an interpreter is then a sort of homunculus . . . Therefore, psychology without homunculi is impossible. But psychology with homunculi is doomed to circularity or infinite regress, so psychology is impossible. (1978, p. 122)[9]

What all of this suggests is the following. If cognitive scientists are going to invoke a notion of representation in their theories of cognition, then although such a posit will need to share some features with our commonsense notions (to be recognizable as representations), the scientific account cannot simply transplant the commonsense notions and leave it at that. The folk notions, as such, are ill-suited for scientific theories because they carry features whose place in the natural order is unspecified. Hence, some further work is needed to account for these features and show how

[9] Dennett argues this dilemma is solved in the CCTC through the "discharging" of the interpreter/homunculus. This involves explaining the sophisticated capacities of the homuculus/interpreter by appealing to increasingly less sophisticated components that comprise it. We will again return to a discussion of this strategy in forthcoming chapters.

representation can be part of a naturalistic, mechanistic explanation. There needs to be some unique role or set of causal relations that warrants our saying some structure or state serves a representational function. These roles and relations should enable us to distinguish the representational from the non-representational and should provide us with conditions that delineate the sort of job representations perform, *qua* representations, in a physical system. I'll refer to the task of specifying such a role as the "*job description challenge*." What we want is a job description that tells us what it is for something to function as a representation in a physical system.

What might a successful job description for cognitive representation look like? Part of this will depend on the particular sort of representational notion invoked in a given account. But there are more general criteria that we can expect to be eventually met whenever a notion of inner representation is put forth as part of a naturalistic theory of cognition. These are conditions that need to be elucidated if the invoking of inner representations is going to do any real explanatory work. In the case of reductive theories – that is, theories that attempt to explain cognition as something else (like computation or neurological processes) – representation cannot simply serve as an explanatory primitive. If we are to understand these processes as representational in nature, we need to be told, in presumably computational, mechanical or causal/physical terms, just how the system employs representational structures. Principally, there needs to be some sort of account of just how the structure's possession of intentional content is (in some way) relevant to what it does in the cognitive system. After all, to be a representation, a state or structure must not only have content, but it must also be the case that this content is in some way pertinent to how it is used. We need, in other words, an account of how it actually *serves as* a representation in a physical system; of how it functions as a representation. Dretske captures exactly the right idea: "The fact that [representations] have a content, the fact that they have a *semantic* character, must be relevant to the kind of effects they produce" (Dretske 1988, p. 80). For the moment, we can leave unspecified exactly what "relevant" means in this context. As we saw in section 1.1.1, specifying the relevancy is tricky business because on several accounts of content, it is far from clear how the content itself can be a *causally* relevant feature of a structure. And if these conditions aren't causally relevant, it is far from clear how they can be explanatorily relevant or how they can be at all "relevant to the kind of effects they produce." For now, we can simply stipulate that the positing of inner representations needs to include some sort of story about how the structure or state in question actually plays a representational role.

Specifying how a posited representation actually serves as a representation is important because representation is, as Pierce and others have emphasized, a functional notion. Without the functional story, it would be virtually impossible to make sense of this aspect of the theory. Consider the following analogy. Suppose someone offers an account of some organic process, and suppose this account posits the existence of a structure that is characterized as a pump. The advocate of the account would need to provide some sort of explanation of how the structure in question actually serves as a pump in the process in question. Without such a story, we would have no reason for thinking that the description is accurate or that there are any structures that actually *are* pumps. Now suppose that when we ask how it is that the structure in question functions as a pump, we are told that it does so by absorbing some chemical compound, and nothing more. In this scenario, we would properly complain that the role the structure is characterized as playing is not the role associated with our ordinary understanding of a pump. To be a pump, an entity must, in some way, transfer material from one place to another. What the theory appears to posit is not a pump, but instead what sounds more like a sponge. Because functioning as a sponge is clearly different than functioning as a pump, then despite the way the theory is advertised, it would belong in the class of sponge-invoking theories, not pump-invoking theories.

In a similar manner, cognitive researchers who invoke a notion of inner representation in their reductive accounts of cognition must provide us with some explanation of how the thing they are positing actually serves as a representation in the system in question. We need to be given a description of the structure that enables us to see how it does something recognizably representational in nature. If we are told that it is a representation by virtue of doing something that no one would think of as a representational role – say, by functioning as a mere causal relay – then we would have good reason to be skeptical about the representational nature of the account. Indeed, if the role described is one that is shared by a wide array of other types of entities and structures, we would have the additional problem of representations (in the dubious sense) appearing everywhere. In other works, I've referred to this as the "problem of pan-representationalism" (Ramsey 1995).[10] A central goal of this book is to argue that this hypothetical situation is in fact the actual situation in a wide range of newer cognitive theories.

[10] Fodor has used the term "pansemanticism" to make a similar point (Fodor 1990).

It is important to recognize that meeting the job description challenge is not the same thing as providing a naturalistic account of content. The latter would present the set of physical or causal conditions that ground the content of the representation – the conditions that determine how a state or structure comes to have intentional content in the first place. A complete and fully naturalistic account of representation would need to provide such an account since without it a central aspect of representation would remain mysterious and unexplained. As Dennett might put it, the "intentionality loan" associated with the positing of a representation would go unpaid (Dennett 1978). As noted above, providing such a set of conditions has been a major project in the philosophy of mind for some time. Some of the more popular attempts to explain content naturalistically are accounts that appeal to types of nomic dependency relations (Dretske 1988, Fodor 1987), causal links to the world (Field 1978), evolutionary function (Millikan 1984), and conceptual roles within a given system (Block 1986). Yet insofar as the goal of these theories is to explain a certain type of *relation* – the relation between a representation and its intentional object – they are not accounts that directly explain what it is for a state or structure to actually *function as* a representation in a physical system. As we will see in forthcoming chapters, it is true that some of the accounts of content also strongly suggest certain strategies for answering the job description challenge. But viewed strictly as accounts of content, that is not their primary objective.

To see this distinction better, consider the various circumstances associated with the normal use of a compass. On the one hand, we might be interested in knowing how a compass actually functions as a representational device. What makes a compass serve as a representational device is the fact that the position of the needle literally serves to inform a cognitive agent about different directions. The content of the compass is relevant to its job because the needle's position is used to reveal facts about, say, the orientation of magnetic north. It is in this way that the compass comes to serve as a representation. There are, of course, many ways it might do this, and thus there are many different types of compasses. For example, a pocket version of a compass, which is simply held in the hand needle-side up, operates in a manner very different from the stationary versions, like those permanently mounted on the dash of vehicle (which may not even use a needle). But with all compasses, we can see how they play a representational role by seeing how they serve to inform people about directional orientation. This is an understanding of the functionality of a compass.

On the other hand, we might instead be interested in knowing what conditions are responsible for the representational content of the compass – something relevant to, but very different from, the compass's functional role. If we wanted to know how a compass came to acquire its intentional content – the conditions that underlie the information it provides – one obvious answer would be to say that the needle of the compass comes to designate magnetic north because its position is nomically dependent upon (or reliably co-varies with) magnetic north. The semantic content of the compass is thereby grounded in a dependency relation between the needle and a certain condition of the world. It is this dependency relation that makes the needle's position entail certain facts about the world, and thereby enables us to use it as a representational device.

The point here is that the account of how the compass serves as a representation is different from the account of the conditions responsible for its representational content. Of course, in one clear sense, the compass is a poor example for our purposes because it serves as a representation only for a full-blown interpreting thinker, and a thinker is the very sort of thing a theory of the mind can't invoke. But the compass illustrates the need to bear in mind that understanding how a state's content is grounded in some set of conditions is not the same thing as understanding how the state actually serves as a representation. Someone could perfectly well understand the nature of the dependency between the needle and magnetic north and yet be completely ignorant about the way in which the compass functions as a representational device. This bears emphasis because writers sometimes treat naturalistic theories of content as though they provide a complete account of cognitive representation. However, a theory of content is only one part of the puzzle. Any theory that invokes representational structures should also include an accounting of how the posit functions as a representation. The latter would include some sort of accounting of how something's status *as* a representation is pertinent to the way the cognitive system performs some cognitive task.

In fact, in cognitive research, the need to answer the job description challenge for a representational posit is far more pressing than the need to provide a naturalistic account of content. To some extent, researchers can leave the explanation of content to the philosophers. If theorists can develop accounts of cognitive processes that posit representations in a way that reveals just how the posit plays a representational role in the system, then the explanation of content can wait. They can say, "Look, I'm not completely sure how state X comes to get the content it has, but in my explanation of cognition, there needs to be a state X that serves as a representation in the following way." So from the standpoint of psychological theory development, the need for an account of content-grounding is not so urgent. However, if a theorist cannot explain the sense in which a representational posit actually *serves as* a representation, or offers an explanation that is grossly inadequate, then the very representational character of the theory is seriously undermined. In fact, we would have no real reason to think the account is actually representational at all, and whatever pretheoretic understanding of representation we possessed would be irrelevant to our understanding of the cognitive account on offer. In short, a representational theory of cognition should provide, at a bare minimum, an explanation of how something serves as a representation in such a way that, at the end of the day, we still have a *representational* theory, instead of a nonrepresentational account of a psychological process.[11]

The crux of the job description challenge, then, is one of steering a course between the Scylla of putting forth conceptions of representation that are too strong (because central aspects of representation are left unexplained) and the Charybdis of positing conceptions of representation that are too weak (because representation is reduced to something nonrepresentational, uninteresting and ubiquitous). In the case of the former, we would simply have the reintroduction of a sophisticated cognitive capacity (the use and interpretation of representations) with no real understanding of how this is done. In the case of the latter, we would have structures that operate in a way that *is* intelligible, but not intelligible as playing a representation role. What we want is an account of how something described as a representation functions *as such* in a computational or biological system. We want an account that allows us to intuitively recognize the processes in question as distinctively representational, and at the same time illuminates how this comes about.

While I think many authors have recognized certain aspects of the job description challenge, the full nature of the challenge has not been adequately appreciated either by researchers who invoke representations in their theories, or by philosophers attempting to explain representation. It might be assumed that those writers who develop "teleo-semantic" accounts of content come the closest to addressing the worry because their accounts are built on the idea that an appeal to proper function is

[11] In some accounts of higher order representational phenomena, such as memory or conceptual representation, "lower order" representations of features are invoked to serve as constituent elements of larger representational structures. In one sense, these feature representations are often introduced by mere stipulation, without any account of what they are doing that *makes* them into representations of features. Yet at the same time, it could be argued that they clearly are serving as representations in the proposed architecture by functioning as representational constituents of some larger representational system; a fuller discussion of this sort of representation is offered in chapter 3.

the key to understanding cognitive representation. Yet because many of these accounts are focused on handling worries associated with the naturalization of the content relation, the functional role of representing is often invoked without being fully addressed. That is, we are often told that a structure's proper functioning as a representation is critical for understanding, say, how the structure can misrepresent, without being told in sufficient detail what proper functioning as a representation amounts to. For example, Millikan (1984) attempts to provide a general framework that applies to both mental *and* non-mental representations. On one reading of her account, to function as a representation is to be "consumed" by an interpreter that treats the state in question as indicating some condition. Following Peirce, this seems reasonable for non-mental representations – cognitive agents, including non-humans, take advantage of signs and signals in various ways. But insofar as the same account is supposed to apply to internal cognitive representations, it is far from clear how we are supposed to make sense of representation consumption *inside* a cognitive system. Even worse, Millikan seems to allow that processes normally assumed to have a non-representational function, such as the flow of adrenalin caused by threatening situations, really are quasi-representational after all (Millikan 1993). Ultimately, just how a state or structure actually serves a distinctive role of representing is left somewhat mysterious on Millikan's account, and she thus fails to directly answer the job-description challenge. Millikan's account of representation is unsatisfying because she leaves the functionality of representation, in a sense, "under-reduced."[12]

Dretske (1988), on the other hand, does provide an explication of what it is for something to function as a representation in purely mechanistic terms. On my view, Dretske raises exactly the right questions and addresses exactly the right issues. But as we will soon see, it is hard to understand why a structure functioning in the manner Dretske describes should be seen as representational at all. Dretske's account runs into trouble because representation is, in a sense, "over-reduced" – that is, it is reduced to a set of conditions and relations that intuitively have nothing to do with representation at all. In chapter 4, this critique of Dretske's theory will be spelled out in more detail.

It might seem that as I've explained things here, the job description challenge is quite literally impossible to answer. Either you reduce

[12] In fairness to Millikan, it should be noted that her account is quite complex and lends itself to different readings, one of which suggests a notion of representation that *does* meet the job description challenge in the manner that will be suggested in chapter 3.

representation to causal–physical conditions or you don't. If you do, then I'll say you have reduced representation to the non-representational, and therefore you have abandoned the notion of representation altogether. If you don't do this, then you've left some aspect of representation unexplained and mysterious. So, it seems, you're damned if you do and damned if you don't. However, I think it is not only possible to meet the job description challenge, but that this has been successfully done with certain theories in the CCTC paradigm. Chapter 3 will be devoted to spelling out exactly how this works. The point here is that some ways of fitting representations into the natural order reveal why it makes sense to view a given structure as a representation, whereas other ways fail to do this. My claim is that the difference between the two roughly corresponds to the division between classical computational theories and the newer, non-classical accounts of cognition.

Before moving on, it is important to be very clear on exactly what answering the job description challenge involves. The challenge involves answering neither of the following questions:

(a) Is it possible to describe physical or computational processes in representational terms?

(b) Is it absolutely necessary to describe physical or computational processes in representational terms?

The shared problem with these two questions is that they lend themselves to trivial answers that are uninformative. Consider question (a). As Dennett (1978) and others have noted, it is indeed possible to adopt the "intentional stance" toward practically every physical thing and system. Even a rock can be described as acting on the belief that it needs to sit very still. So, trivially, it is always *possible* to characterize physical systems using representational language by adopting this perspective. However, we tend to find this sort of intentional characterization gratuitous and unnecessary, in part because the notion of representation involved fails to be sufficiently robust. Showing that a system is representational in *this* sense isn't terribly informative in helping us to understand what might be going on with representational cognitive systems.[13]

Going the other way, it is always possible to describe a physical system using purely non-intentional, causal/physical terms. Just as we can avoid biological language in the description of biological systems by dropping

[13] Dennett himself would disagree, as he believes it is a mistake to try to characterize mental representation as a concrete state playing a certain role. Instead, he believes the distinction between representational and non-representational systems is entirely a function of the usefulness of adopting the intentional stance. For Dennett, the whole enterprise of trying to understand representation in the manner described here is misguided.

down to the level of molecules and atoms, so too, it will never be absolutely *necessary* to invoke representational language in the characterization of a representational system. So the answer to (b) is trivially "No." Invoking Dennett's terminology once again, it is always in principle possible to adopt the "physical stance" (using only the terms of basic physics) toward any physical system, even when robust and sophisticated representations are working within the system.

Hence, the job description challenge requires us to address questions that are more nuanced than these. The sorts of questions that need answering are more along the lines of the following:

(c) Is there some explanatory benefit in describing an internal element of a physical or computational process in representational terms.

Or maybe:

(d) Is there an element of a proposed process or architecture that is functioning as a representation in a sufficiently robust or recognizable manner, and if so, how does it do this?

Or even:

(e) Given that theory X invokes internal representations in its account of process Y, are the internal states actually playing this sort of role, and if so, how?

Now unfortunately, neither (c) nor (d) nor (e) is as clean or crisp as we would like. Exactly what counts as representations having an "explanatory benefit" in a theory, or just what it means to be a "sufficiently robust" notion of representation – these are vague matters that, to some degree, require our making a judgment call. But as noted earlier, this shouldn't be terribly surprising. We want to know if, given what a theory says, something is actually functioning in a manner that is properly described as representational in nature. We can't do this without making use of our ordinary understanding of representation and representation function. And, as with any form of concept application, this requires a judgment call. So be it. As we proceed with our analysis, the outlines of what is and is not involved in a "sufficiently robust" notion of representation having an "explanatory benefit" will start becoming increasingly clear.

1.3 DEMARCATING TYPES OF REPRESENTATION AND TYPES OF REPRESENTATIONAL THEORIES

Since we are going to explore different notions of representation in cognitive science with regard to how well they meet the job description challenge, more needs to said about the demarcation strategy I plan to use to classify different notions of representation. There are, of course, a variety of different ways we can distinguish representational notions. The most popular strategies appeal to either the way in which the representation is thought to acquire its content (e.g., nomic dependency vs. conceptual role), or the form or structure of the representation (e.g., compositional form vs. distributed representation), or the sort of theory in which it appears (the classical computational vs. connectionist models).[14]

While all of these taxonomies have their advantages, I am going to use a scheme that is better suited for the issues we will be addressing. In what follows, we need to demarcate notions in a way that places the explanatory role of representations at center stage. My taxonomy will individuate types of representations in terms of the conditions thought to be constitutive of the representational role. That is, I'll be grouping together representational posits from various theories if they all appeal to the same or similar factors to justify the claim that something is serving as a representation. For example, one representational notion we will explore is based on the idea that structures function as elements of a model or simulation of some target domain. Another notion stems from the idea that something functions as a representation because it is reliably triggered in a certain way. Carving things up this way will do two things for us. First, we can avoid examining each individual theory in cognitive science because we can cluster theories together that employ the same basic representational ideas. Second, it will focus attention on what I take to be the most critical feature of representational posits. By making the criteria for type-identity those factors in virtue of which something is claimed to serve as a representation, we will direct the spotlight on the aspect that matters the most for our interests.

Another issue that needs clarifying concerns the way representation is linked to the overarching explanatory goals of a cognitive theory. I've been treating notions of representation as a type of theoretical posit, put forth as part of an explanatory apparatus to account for some cognitive ability or process. On this construal, notions of representations are *explanantia*. However, a great deal of work in cognitive science is also devoted to explaining the nature of representation itself. Various accounts of knowledge, memory, imagery, concepts and other intentional aspects of the mind take representation to be the very cognitive phenomenon that one is attempting to explain. On this construal, cognitive representations are the *explanandum* of a given theory. How does my analysis bear on theories that don't posit representational states but rather try to explain them?

[14] See, for example, Fodor (1985).

First, the distinction between theories that posit representations and theories that try to explain representation is not as sharp as one might assume. A large number of cognitive models – indeed, perhaps the majority – do a little of both. For example, production systems like Allen Newell's SOAR model can be seen as an attempt to explain how humans do certain types of problem solving by providing an account of a specific style of knowledge representation and retrieval (Newell 1990). Newell's theory employs a computational framework to explain both general cognition and the way we represent problems and their solutions. Moreover, we've seen that reductive theories that posit representations (as explanantia) also offer a story about the way representations are implemented in the structures or states of their overarching architecture. This part of the account works as a mini-theory about the nature of representation itself. Thus, the division between theories that treat representation as explanantia and those that treat them as explananda is not very sharp.

Second, it seems clear that meeting the job description challenge should be a goal for theories of representation every bit as much as it is for theories of cognition that invoke representations. Surely any reductive account that is designed to explain the nature of mental representation, or a specific sort of mental representation, will need to spell out how the system employs the state or structure *as* a representation. For example, if the theory is a reductive account of memory that attempts to explain how we store and retrieve representations of long-term knowledge, then it will need to be shown how the parts of the system allegedly responsible for this really do function as representations with intentional content. Answering the job description challenge should be a primary goal of any theory of representation. Hence, much of what follows will be relevant to those reductive accounts whose primary goal is to explain cognitive capacities by appealing to representations, *and* those whose main objective is to explain mental representation itself.

1.4 SUMMARY

In this chapter, the aim has been to introduce some of the issues and concerns that will occupy us in the subsequent chapters. We've seen some of the intuitive aspects of our commonsense understanding of representation that can and do become incorporated into the more theoretical notions found in cognitive science. We've also seen some of the problems associated with those intuitive aspects. One problem is determining how representational content fits into the natural world. But cognitive theories that invoke representations carry the greater burden of providing an adequate account not of content, but of representational function – of how something serves as a representation in the proposed architecture. This is the job description challenge, and we will be returning to it throughout our analysis of different conceptions of representation. The challenge is to explain how a given state actually serves as a representation in a way that is both naturalistically respectable and doesn't make representation completely uninteresting and divorced from our ordinary understanding of what representations are.

In what follows, we will see that some theories employ notions of representation that lend themselves to a successful assimilation into the natural order while retaining their status as real representations. Other theories employ notions of representation that do not successfully assimilate. By and large, the successful notions are found in the CCTC while the unsuccessful notions can be found in connectionist models or the various accounts in the cognitive neurosciences. So I'll be offering both a positive and a negative account of representation in the cognitive sciences. However, before we can start comparing these different accounts, we first need to remove what I take to be an ill-conceived interpretation of the notion of representation in the CCTC. We saw above that naturalistic theories cannot just co-opt the folk notions of mental representation when constructing their accounts of cognition because the folk notions come with features whose place in the natural order is unclear. Further explication is required. But another danger is that the intuitive nature of our commonsense framework will permeate and cloud our understanding of what might actually be a more technical notion of representation that is not directly based on folk psychology. In large measure, I believe this has happened with our current understanding of the CCTC, resulting in a mistaken interpretation of how the CCTC is committed to inner representations. Showing this will be the goal of the next chapter.