Chapter 8

Philosophy: Foundations of Cognitive Science

8.1 Philosophy in Cognitive Science

All sciences used to be branches of philosophy. A science is born when it breaks off from philosophy and begins to be pursued by specialists. Physics, biology, and chemistry are all sciences that were born in this way a rather long time ago, but all began as branches of philosophy. Cognitive science, and the disciplines it comprises—psychology, linguistics, neuroscience, and computer science (omitting, for the moment, philosophy itself)—are young sciences, each having emerged from philosophy within the last hundred years or so. Psychology was born as an independent science in the last few decades of the nineteenth century. Neuroscience had its beginnings around the same time, though its real development into a promising theoretical enterprise is much more recent. Linguistics as we know it today began to emerge in the 1920s and is still very much in the process of becoming independent, with certain of its problems, particularly those having to do with logic and semantics, still falling as much in the domain of philosophy as in its own. Computer science has existed only since about 1950 (though its roots are primarily in mathematics).

Historical Background

Cognitive science, the fusion of these disciplines, is younger still, only several decades old. So it, even more than its component disciplines, is thoroughly entwined with its philosophical roots. These roots are to be found in the seventeenth century, when philosophers began to find new ways of addressing problems about the nature of thought and the mind. Debates began about the relation between mind and body, the relation between language and thought, the relation between thoughts or perceptions and the objects thought about or perceived, whether ideas are innate or acquired, and the nature of the embodiment of mind. Amid this intellectual ferment, two figures stand out as grandfathers of the cognitive approach: René Descartes and Thomas Hobbes.

Descartes argued that all of our knowledge of the external world is mediated by *representations*—mental objects that somehow stand for things outside. Thought, he contended, always involves the manipulation, through inference or other mental processes, of these representations. This is not as obvious as it might seem at first. After all, it could be (and many have argued that it is the case) that our knowledge of the world consists merely of our being able to do certain things and that it in no way involves manipulating internal symbols (whether they are made of immaterial soul-stuff as Descartes thought, or of gray matter, or of silicon). One natural way to think about these representations is that adopted by many of Descartes's contemporaries—as mental images of what they represent. Another, more plausible, and ultimately more

influential way is as sentences (or as remarks of some kind) in an internal language of thought, or perhaps in the native language of the speaker. What makes the contention that thought is representational so interesting are the implications Descartes saw in this position, implications that have helped to shape not only all subsequent philosophical thought about the mind but current cognitive science as well.

The first implication Descartes noticed was that these representations have no necessary connection to the things they represent. Of course, fortunately for us (and, Descartes thought, only through the good graces of the deity), they tend to bear some consistent relation to the things they represent and hence are fairly well able to serve their function of guiding our activity in the real world. But, Descartes suggested, even if there were no external world at all, we could have the same representations that we do now (just as, though there are in fact no unicorns, there are pictures of unicorns, and just as, though we have pictures and reported sightings of yetis, we don't know whether these yeti representations correspond to actual yetis). Of course, we would be mistaken in thinking that they represented reality, but they would be the same mental states. They would feel the same, would interact with each other in the same way, and would guide behavior in the same way. (Of course, who is to say that we are not in that state right now?) Because of the skeptical worry that this position sometimes raises, let us for now call Descartes's insight *representational skepticism.*

The second interesting implication of Descartes's view is that one can study the mind without paying any attention at all to the reality it purports to represent and think about. After all, on this view, since what we are studying is just the nature and interrelations of symbols and processes going on inside a mind, and since those symbols and processes would be what they are even if nothing but that mind existed, why bother paying attention to anything but those symbols and processes themselves? The suggestion is not that nothing other than the mind in fact exists—a view called *solipsism*—but rather that if we study the mind as if solipsism were true, we can say everything scientifically interesting that we would ever want to say about it. For this reason, the view is called *methodological solipsism.* (For more discussion of methodological solipsism, see Putnam 1975b or Fodor 1981, chap. 3.)

The third interesting inference that Descartes drew from his *representational theory of mind* is that mind and body are two completely different kinds of thing. This view, for which Descartes is perhaps best known, is usually called *Cartesian dualism.* Most often dualism is interpreted as the belief in a ghostly soul-stuff permeating our bodies in some mysterious (nonspatial) way and running them for us. This is probably roughly how Descartes thought of it. But we can think of Cartesian dualism in a slightly more sophisticated light: mental things, like beliefs, images, and thoughts, are what they are because of what they represent. This, after all, is the central insight of the representational theory of the mind. Now, they represent what they represent because of how they behave in the mind (after all, we just saw that it can't be because of any relation they bear to the external world). The point is that very different kinds of things (brain states, states of computers, ink marks on paper, sounds) could all represent the same thing. For instance, consider the situation depicted in figure 8.1.

The figure shows John, John's name (*John*), the image of John in Bill's head, Bill's thought about John, the sounds Bill makes when he calls John, Bill's picture of John (in whatever style Bill is working in these days), and a magnetic record of 'John' in a computer. Except for John, each of these things is a representation of John. But these representations have nothing whatever in common physically. What makes them

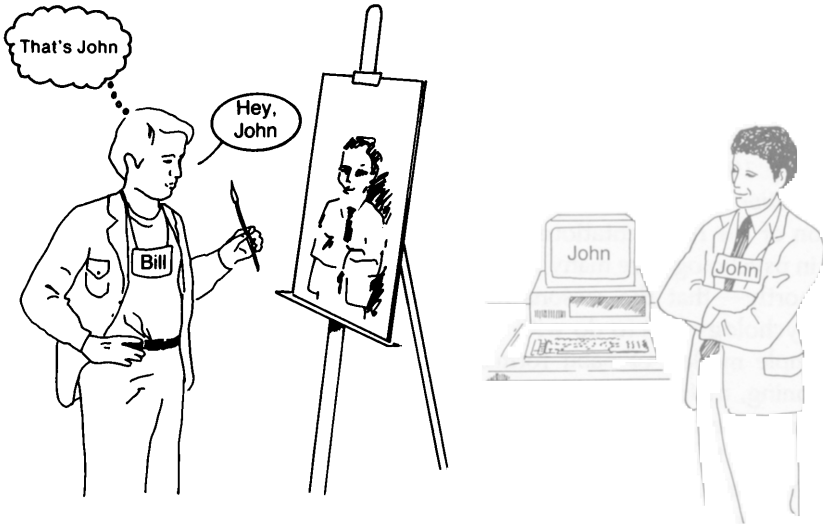

Figure 8.1
John and some representations of John

about John must therefore be, a modern-day Cartesian can reason, something nonphysical. Hence, what makes a representation the representation that it is, this line of reasoning continues, is some nonphysical fact about it. Mental objects, considered as mental, are therefore nonphysical kinds of things. This is not to say that representations are not also physical things. After all, the image in Bill's head, the painting, the name tag, the sounds, and the computer record are all physical, but the kind of thing they all are—a representation of John—is a nonphysical kind of thing.

These three central tenets of the representational theory of mind—representational skepticism, methodological solipsism, and Cartesian dualism—have been extremely influential in the history of thought about the mind. All three tenets are represented to some degree in contemporary cognitive science. They are Descartes's legacy.

Hobbes introduced one twist on Descartes's view that is interesting for our purposes. (Do not be misled—Hobbes and Descartes had very different views of the mind and differed from one another radically on many points, but this particular one of Hobbes's insights can usefully be grafted onto the Cartesian theory we have just outlined to yield an intriguing picture.) Hobbes suggested that "all reasoning is but reckoning." By this he meant that thought can be understood as a kind of calculation, perhaps often unconscious, using formal operations on symbols stored in the mind. With Hobbes's dictum in mind, we can see the completion of Descartes's model of mind as a prototype for contemporary cognitive science. Not only are our mental states and processes to be conceived of as forming a sort of autonomous representational system; in fact, they are all to be thought of as in some sense *mathematical* (or at least *linguistic*—but at any rate, in some sense *formal*) objects, at least at some level of description, and the operations our minds perform on them when we think are to be conceived of as *computations*. Again, note that this elaboration is neither obvious nor in any sense necessarily true. It could well be (as many current connectionists maintain) that although we represent the world, our representations are not in any appropriate sense distinct formal objects themselves. These fruits of seventeenth-century philosophy contain the seeds of contemporary cognitive science.

Over the next three hundred years the Cartesian approach to the philosophy of mind went in and out of fashion and was refined and blended with other approaches. At the end of the nineteenth century philosophy gave birth to psychology. That first psychology, sometimes called *introspectionism*, was very Cartesian in its orientation, but it soon gave way to *behaviorism*, a decidedly anti-Cartesian school of psychology. Then, in the early 1950s, a remarkable development occurred. Behaviorism began to give way to *cognitive psychology*, a brand of psychology that takes seriously the computational version of the representational theory of mind. (Though the reasons for this development in psychology are many and complex, one can say with some justice—as we will see shortly—that the reasons for the demise of behaviorism and the rise of cognitive psychology had to do with the difficulties behaviorists had in extending their rather simple models of habit formation and learning to theories of complex behavior, reasoning, memory, problem solving, language acquisition, and the like—exactly the things Descartes and Hobbes found most interesting. The patterns of failure suggested that theories of a very different kind—describing internal representational structures in detail (or perhaps brain processes)—would be necessary to account for this range of phenomena.) At the same time philosophy began to swing back in the Cartesian-Hobbesian direction, linguistics (in Chomsky's very Cartesian form) began to emerge as an exciting science, and computer science emerged as a full-fledged discipline. Motivated by that same Cartesian-Hobbesian vision of the mind as a calculating device operating on representations, computer scientists began the quest for artificial intelligence. Cognitive science was conceived.

The Role of Philosophy

Clearly, philosophy has played an important role in the history of cognitive science and in the history of the ideas it embodies. The philosopher also has a place in the ongoing practice of cognitive science. Philosophy is a foundational discipline. Not only does it do the spadework that makes the construction of other disciplines possible; it also pays constant attention to the foundations of those disciplines as they are practiced. Philosophers assist scientists in defining their enterprise and in clarifying what they are studying, what their methods ought to be, and what relations hold between the entities studied by the various disciplines. This function is particularly important in a new, interdisciplinary enterprise like cognitive science, in which the entities being studied—abstract mental and computational processes—are often difficult to pin down and in which practitioners of different disciplines are working on related problems in rather different ways. The philosopher helps these collaborators to formulate their problems and models and to think more clearly about the nature and structure of the objects and processes under discussion. Philosophers have worried about these questions about the mind and language for a few millennia, and, if nothing else, they know where it is easy to get muddled. We can distinguish three areas of philosophical contribution: defining the enterprise and getting a synoptic view of it (*philosophy of science*); concerning itself with the nature of the abstract structures being studied by cognitive science, and their relation to more concrete things (*metaphysics*); and thinking about the interrelations between representations, and how the mind organizes and uses them to generate knowledge (*epistemology*).

In the following sections of this chapter we will apply all of these philosophical contributions: we will attempt to gain a synoptic view of the enterprise of cognitive science, consider ontological questions that it raises, discuss knowledge (how it is

represented in the mind and how, if at all, it could be represented in a machine), and explore the current state of the field.

8.2 *The Enterprise of Cognitive Science*

Behaviorism

The first fifty years of this century saw the study of the mind dominated by a school of psychological thought known as *behaviorism*, led by I. P. Pavlov, John Watson, Edward Chace Tolman, Clark Hull, and B. F. Skinner. Behaviorism arose as a reaction against introspectionism. The introspectionists studied the contents and structure of consciousness by having carefully trained subjects introspect, that is, "look inside" their minds and report what they observed, under carefully controlled conditions and while performing particular cognitive tasks. Introspectionist psychology failed largely because of the fallibility of introspection, which was, after all, its principal instrument. But the features of introspectionist psychology most directly responsible for its collapse were the tremendous disagreements over fundamental data between different laboratories and the lack of any unified, testable theory to explain these data.

The behaviorists argued that the problem lay in the subjectivity of the introspectionist method. They contrasted this method, whose data could be directly observed only by the subject and could not be independently verified, with the methods of the physical and biological sciences, where data are always public and therefore independently observable, or objective. Some behaviorist critics went so far as to suggest that the very phenomena the introspectionists claimed to be studying—the mind, consciousness, attention, and cognitive processes, among others—could not even be shown to exist and were therefore not proper objects of scientific inquiry at all. The behaviorists proposed to replace introspectionism with an objective science of behavior modeled on the more successful physical sciences.

Behaviorism was not confined to psychology. Philosophers such as Gilbert Ryle, Ludwig Wittgenstein, Rudolph Carnap, Otto Neurath, and Moritz Schlick argued that if they were to be at all useful to a science of psychology, such *mentalistic* terms as *thought*, *belief*, *mind*, and *consciousness* had to be redefined in terms of, or replaced with, more objective terms that referred only to publicly observable movements of the organism or to events in its environment.

The behaviorists attempted to discover *scientific laws*, that is, universal generalizations that would describe, predict, and explain the relations between the *stimuli* organisms encountered in their environment and the *responses*, or movements, they produced in the presence of those stimuli. The principal area investigated by the behaviorists was learning. Laws were sought, for instance, that would predict the rate at which rats would press bars when they received food rewards on a variable rather than a fixed schedule.

As an approach to explaining simple sorts of behavior, particularly of cognitively simple animals in carefully restricted situations, behaviorism was rather successful. But when the behaviorists attempted to understand more complex behavior, they encountered difficulties. It became apparent that there simply are no good ways to describe very complex behavior such as speech that allow the formulation of explanatory laws. For example, try to develop a generalization that enables you to predict under what circumstances someone will use an adverb, just as a physicist can predict the orbit of a

planet. Moreover, much behavior (again, linguistic behavior provides an excellent example) does not seem to be under the direct, lawlike control either of stimuli in the environment or of past reinforcement or punishment histories. Rather, it seems to be generated by complex cognitive structures, including those responsible for the thought being expressed and such structures as the grammars studied by linguists.

Behaviorism claimed that all mentalistic terms can be redefined in terms of observable, physically describable behavior. But consider any mentalistic term—say, "thinking that the queen of England is the richest woman in the world." A behavioral definition of that term might run like this: "'Thinking that the queen of England is the richest woman in the world' $=_{df}$ (abbreviation for "is by definition") *being disposed to say things like 'The queen of England is the richest woman in the world'; being disposed to answer the question 'Who is the richest woman in the world?' with 'The queen of England, of course'; and so on.*" (Definitions like this, with many variations, have been suggested by defenders of various versions of behaviorism.)

But such definitions cannot work, for several reasons. First, the *and so on* at the end of the definition is not just an abbreviation for a lot of dispositions that we are just too lazy to specify. The number of dispositions necessary to fill out such a definition is boundless, and even if endless definitions make sense, they are certainly of little use to a scientific discipline.

Second, many things that do not have the requisite beliefs nonetheless have the named dispositions—for example, the tape recorder with the tape loop that endlessly plays "The queen of England is the richest woman in the world."

Third, there is the case of things that do have the belief in question but lack any of the supposedly defining behavioral dispositions. Suppose you are being tortured by the Renganese secret police, who want to know who the richest woman in the world is so that they can kidnap her, hold her for ransom, and solve their national debt problems. You know the answer. But do you have the disposition to answer their questions correctly? Certainly not! (And assuming you did crack under torture, for a die-hard behaviorist, your answer could be of no use to your captors. Although their interpreter might come to say things like "The queen of England is the richest woman in the world," the Renganese secret police speak only Renganese. So they would only come to say things like "Hoya pata Englaterri nyool chen mikya"—which is a different thing to say and hence bespeaks a different belief, on the standard behaviorist account.)

Fourth, a behaviorist might try to rescue the account by saying that the dispositions are dispositions to say things only when a speaker *wants* to evidence the belief, or when the speaker *believes* that no harm will result, or some such thing, and that not only English words count, but so do any translations of them, that is, any words that *mean the same thing*. But both of these attempts, though perhaps the only hope for saving the theory, lead down the garden path to circularity. The goal of behaviorism was to define mentalistic terms using only behavioral terms, but this strategy for rescuing failed definitions relies upon using mentalistic terms themselves to define other mentalistic terms. Thus, behaviorism appears to succumb not only to empirical difficulties, but to conceptual confusion as well.

The failure of behaviorism provided one motivation for adopting the cognitive approach. Behaviorism's difficulties made it clear that in order to really understand complex cognitive capacities, it is necessary to look inside the organism—to pay attention not only to the stimuli impinging upon the organism and its responses to them (though these are certainly important) but also to the internal processes that

mediate between perception and action. But what was lacking until recently, when computer science developed, was a suitable model for the internal processing that could support such behavior.

The Computer and Cognitive Science

The necessary model for internal processing was supplied by the digital computer. Computer science showed cognitive scientists that it was possible to explain the intelligent behavior of a complex system without presupposing the intelligence of its components by employing the idea of an information-processing system and the computational model of explanation. This model also demonstrated the possibility of analyzing meaning in terms of *functionally interpreted* states of physical systems (what Newell and Simon (1976) have called the idea of a *physical symbol system* introduced in chapter 2). Finally, all of this suggested a model of the relation between mind and body that is respectably *physicalistic*, in that it does not posit a dualism of substance, but that avoids the pitfalls of behaviorism and does not involve reducing the mental to the physical. Let us first see how the digital computer gives r se tb each of these ideas, and then turn to their realization in the enterprise of cognitive science.

For many years, especially in the heyday of behaviorism, it was thought that there were only two ways to explain intelligent behavior: physicalistically or mentalistically. Mentalistic explanations operated by reference to the "internal workings of the mind" or, as Ryle (1949) called it, the "ghost in the machine." Such explanations were looked upon with disfavor because it was argued (by Ryle, among others) that any mentalistic explanation could only "explain" intelligence by appealing to structures or processes that were themselves intelligent.

As an example, consider a hypothetical debate between yourself, as a cognitive scientist, and a philosopher of Ryle's school. You want to explain your ability to come up with examples in a discussion of philosophy of mind by appealing to internal processes. "Well," the philosopher might say, "come up with some plausible candidate processes (not necessarily all of the details)." You might reply, "There is a process that selects important features of the topic under discussion, and a 'librarian process' that checks my memory store of philosophical examples for examples that have some of those features, and a 'pattern matcher' that finds the closest one to what I need, and an 'augmenter' that fixes up the details just right."

With only this much in hand, the philosopher can argue, "These internal processes are all well and good, but each must be applied intelligently. If the 'feature selector' is to do any good, it must select the right features; if the 'librarian' is to do its job well it must select the best examples; and so forth. And all of these tasks require intelligence. Therefore, your explanation of your ability requires us to explain the intelligence of your subsystems, and we are back where we started, only with more problems. So much for ghosts in machines."

The advent of digital computers has provided a model of explanation that suggests a reply to this argument (which we may call *Ryle's regress*). Imagine how we explain the ability of a computer to do the things it does. We posit subprocesses to explain the actions of processes, sub-subprocesses to explain the actions of subprocesses, and so on, until we reach the level of elementary information processes. Although the action of the whole program may appear to require brilliance (especially if it reliably generates good philosophical examples), the processes into which it decomposes at the first level (the "main" subroutines) require only moderate brightness. As we go down

through the *levels of decomposition* in the explanation, the spark of intelligence required for the processes at each level gradually dims, until we reach the machine language instructions, which are easy to implement mechanically. The ghost is exorcized by gradually reducing it to simple formal operations as we elaborate the explanation.

From the idea of an information-processing system it is but a short step to the idea of a physical symbol system. We will make both of these concepts much more precise later, but for now note that what a computer does is process *symbols*. Symbols always have a dual nature. On the one hand, they are physical things (like the ink on this page, or the electrical impulses and magnetic records in the computer); on the other hand, they stand for things other than themselves. The computer processes these symbols according to rules, and the meanings of the symbols are tied up with these rules. But the computer does not need to "know" the meanings of the symbols. It performs its operations on symbols by means of procedures that depend only upon their physical characteristics. The trick, of course, is to get the physical and meaningful (or semantic) characteristics of the symbols, the rules, and the machine employing them to match up in the right way. This is the essence of an information-processing system: that it encodes information about the world, operates on that information in some way that can be characterized as meaningful, and is structured as a set of functionally organized, interacting parts. The digital computer is a perfect example of these ideas, and it has provided dramatic evidence that intelligent performance can be the product of a physical symbol system.

The final item in our catalogue of ideas given to cognitive science by the digital computer is an account of the relation between mind and body that is neither objectionably dualistic (as in a naive Cartesian theory of mind) nor objectionably reductionistic (as in a naive behaviorist theory of mind). The idea, known as *functionalism*, is that mental states, such as beliefs, and mental processes, such as considering or deciding, are nothing but physical states described functionally. The same physical state in differently organized systems might yield different mental states; the same mental state might be realized very differently in different physical systems.

This is the case with computers. When a small personal computer and the largest supercomputer perform the same computation, they have little in common from a physical standpoint, though functionally they may be identical. Similarly, a computer performing a particular computation using a particular machine language subroutine may in one program be deciding on a chess move (if that subroutine in that context is a "position evaluator"), and it may in another program be deciding whether to buy pork bellies (if in that context it is an "expected gain estimator"). The low-level computational states (and so, perhaps, the physical states) are the same, but the high-level functional interpretations they receive are radically different.

This fact about computers certainly suggests an intriguing question: Might the same be true with human minds? Our psychological states might not be reducible to our physical states, since different physical states might be correlated with the same psychological state, and vice versa. But this does not entail a dualism of substance. Each particular (or *token*) psychological state is some particular physical state (a view known as *token identity theory*), but no *type* of psychological state is a type of physical state, and vice versa. This means that there need be no mystery about what kind of thing any particular psychological state or process is—it is some particular physical state or process—but we need not be committed to the view that whenever that physical state or process occurs in a person, the same psychological state or process is occurring in that person, or vice versa.

These are the ways in which the digital computer facilitated the transition from behaviorism to cognitive science. Let us now turn to the questions the cognitive scientist asks of the philosopher of science. What is cognitive science's characteristic method of asking, attacking, and answering questions? How exactly does it conceive of the mind? We begin with the idea of an information-processing system.

Information-Processing Systems

What is an information-processing system? This question goes right to the heart of the structure of cognitive science, since more than anything else, the view of the mind as an information-processing system is what characterizes and unifies the field. We discussed the general concept of an information-processing system in chapters 1 and 2, and we will introduce other basic distinctions here in order to set the stage for the rest of our discussion.

The first distinction is between *digital* and *analog* representations. Although the digital computer is the dominant technology in present-day computing, there is another type of electronic computer, which is based on analog computation. A mundane example will illustrate the characteristics of these analog computers. Most people buy things using digital monetary systems that are restricted to fixed denominations: dollars and cents, pounds and shillings, and so on. But in some places people use analog money systems in which goods are exchanged for quantities of some material, say, gold or silver. Gold has a certain advantage over dollars and cents: the value of a piece of gold is directly proportional to its weight, an intrinsic physical characteristic. In digital money systems, such as the U.S. system, no such simple relationship holds between the values of coins and bills and any of their physical properties. That is, there is no function relating value to weight, size, or any other basic physical characteristic. Instead, for each type of bill and coin there is a rather arbitrary relation between some cluster of physical characteristics and a particular fixed value. A penny is identified by a set of physical characteristics, but there is no law of physics that will allow us to predict the characteristics of the nickel and dime from the characteristics of the penny. Further, this kind of arbitrary mapping requires that some coin represent the smallest value in the system.

A nice consequence of the direct physical relation used for gold is that it can be used to represent *continuous* monetary values, because weight varies continuously. Weight, and therefore gold, also has no minimum value (we are ignoring questions that arise at the atomic scale). Suppose you have some particularly worthless object, say, an outdated textbook, good for nothing but holding up table legs. Suppose you want to sell it to a neighbor with a rickety table for .7 cents. A transaction in dollars and cents is impossible, but gold meets your requirements. If gold is worth $200 an ounce, then .0000035 ounces of gold is exactly what your neighbor owes you for the book. The analog system appears to be more natural, precise, and flexible.

There is a problem, however. In any computational system we not only need representations; we also need to be able to process them. The apparent precision of the analog system depends on the accuracy of the device used to measure the relevant physical quantity. In order to measure the gold needed to purchase your book, your neighbor's scale would have to be accurate to the ten-millionth of an ounce. This problem is not restricted to attempts to weigh tiny amounts. It arises whenever we need precision that exceeds the capacity of the scale. If the scale is accurate to the hundreth of an ounce, then its range of error at $200 per ounce is $2. Under these

conditions dollars and cents are more accurate (as long as the eyesight of the person handling the money is good enough to distinguish the coins).

Analog systems have advantages. They have a certain simplicity and directness because meaningful values are directly represented as physical quantities. Straightforward physical processes that operate on the physical qualities can therefore have an immediate interpretation. Continuous values can be represented. Also, certain computational problems that are solvable by analog computation lead to unmanageable combinatorial explosions when approached digitally. An example due to Dreyfus (1979) illustrates this point. Suppose you have a map of a complicated railway system linking all of the towns in a region, and you want to figure out the shortest route between two towns. You could write a digital computer program to solve the problem, but any program you wrote could be shown to consume rapidly increasing and finally impractical amounts of time as the number of cities served by the system increased. Instead, you could make a "string-net map" of the system, by tying pieces of string of lengths proportional to the rail links together in such a way that the pattern of links and towns matches the pattern of string pieces and knots. Then you could just grasp the two knots matching the two towns and pull. The taut string path would represent the shortest rail route.

Analog systems also have disadvantages. They always have a margin of error determined by the accuracy of their measuring devices. If values must be held over time, or passed along from process to process, the errors can accumulate rapidly and render the output useless. Because meaningful values and transformations are directly represented by simple continuous physical characteristics and processes, analog systems also tend to be inflexible special-purpose devices. A string-net map is not the right device to schedule trains and keep track of passenger reservations, for example. Thus, it is not clear how a complex and flexible symbol system could be implemented on a strictly analog computer.

The advantage of the digital system is that there is no margin of error. The quantities it represents are always precise. Even a worn penny and a torn dollar bill represent one dollar and one cent exactly. In digital computers symbols are also assigned to physical values in such a way that perturbations of physical qualities rarely cause error. Digital computation, however, often seems awkward and unnatural because of the indirect relationship between complex symbolic structures and physical operations.

Several factors determine what kind of system of representation is appropriate for any particular task. One important variable is the kind of equipment available. If you have good mints and poor scales, choose a digital monetary system; given good scales but poor mints, an analog system might work better. Another important variable is the nature of the task to be accomplished.

An information-processing system may represent and encode information in either a digital or an analog form. If it is to operate digitally, using such perfectly precise, error-free (and hence information-preserving) operations as arithmetic, linguistic, and logical operations, the information must be encoded digitally. If the system is to operate analogically, exploiting the speed and fine-grained nature of such analog processes as rotation, expansion, or continuous amplification, the representations it employs must be analogical.

Just which, if any, human information processes are digital, and which, if any, are analog, is a fascinating philosophical and psychological question. The controversy that most often raises this question concerns the reality of mental imagery. The dispute is

generally couched in these terms: Are all mental representations linguistic in form, or are some pictorial, operated on by processes that can be characterized with terms such as "mental rotation" or "scanning with the 'mind's eye' "? Linguistic representations, of course, are digital. The units of representation are the set of phonemes, or the lexicon of the language, depending on the level of analysis one chooses. The operations on them are the operations of logic, arithmetic, and syntax. Mental images, however, are analog. The units of representation are "pictures" in "mental space," and their dimensions vary continuously. The operations on them are also described spatially, and are continuous. Whether humans use both types of representation in thinking or only one (and if so, which) is an open question in cognitive science. In chapter 2 we discussed the theory that an intrinsically geometric image representation and associated processes are built into human biology (in addition to the references cited there, a good philosophical source is Block 1980c). According to such theories, continuous quantities such as space, size, and angle of view are built into image representation and are transformed by built-in operations for scanning, zooming, rotating, and so on. Although analog imagery theories have been hotly disputed, some theorists believe that they point the way to the discovery of a number of special-purpose analog subsystems in the human mind.

Research on connectionist networks provides a new context for the contrast between analog and digital computation. As we saw in chapter 2, the connection weights, net inputs, and activation values in networks typically vary continuously. Vectors of activation values or weights can often be interpreted as points in multidimensional spaces. Values are computed numerically using continuous transformations, such as the logistic function. Such features suggest that connectionist networks can be thought of as analog computers. On the other hand, the inputs and outputs of connectionist networks often are not numerical analogs of physical quantities. The XOR network described in chapter 2, for example, learns to compute a digital function, although it employs an analog style of computation. Thus, the sense in which connectionism represents an alternative to digital computation is an open question. Later we will return to the issues posed by competing cognitive architectures.

Whether an information-processing system is digital or analog, it is *intentional*. "Intentionality" and its cognates are philosophers' terms for *aboutness*. The thought that the full moon is beautiful is intentional, because it is about, or contains, the full moon. The words in this book are also intentional: they have content; they are about cognitive science. Insofar as the information contained in and processed by an information-processing system is about anything—that is, insofar as it functions representationally—the states and processes of that system are intentional.

Just what it takes for a system or a state of a system to be intentional, and just what it is to represent one thing rather than another, are difficult matters. As noted in chapter 1, it seems at least necessary for there to be an *isomorphism* (sameness of structure) between the representational components of the system and the contents of those representations and processes. Ideally, there would be some structure-preserving mapping between the components of the system that do the representing and the things in the world (or out of it) that they represent. Thus, if all pointers are sporting dogs, and all sporting dogs are carnivorous, then for a person (or a computer database) to represent this set of relations, a structure such as figure 8.2 should be present. The information represented in semantic nets is carried by the structure of the nodes and links. The reason that this net can represent the information it does is that the relations

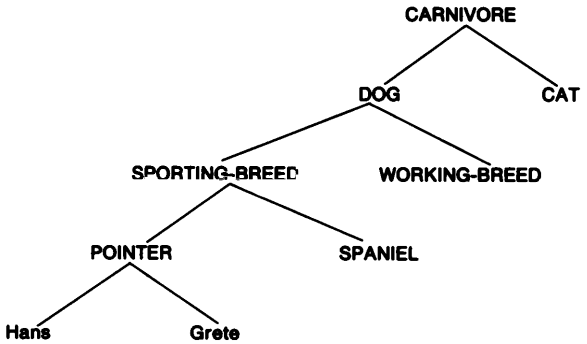

Figure 8.2
Dog hierarchy

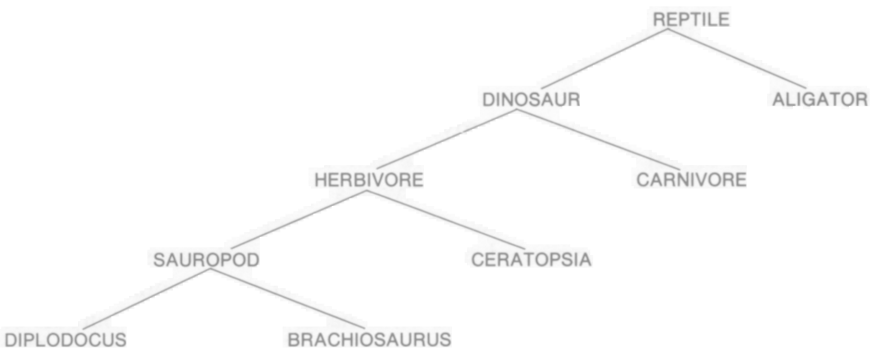

Figure 8.3
Dinosaur hierarchy

between the nodes in the net are isomorphic to the relations between the corresponding entities and classes of entities in the world.

But this relation of isomorphism is not sufficient to make the net *about* dogs and related matters. Consider the net shown in figure 8.3. This net is isomorphic to the one represented in figure 8.2. But even though they are each about something, they are not about the same things. What would capture this difference? One answer is that the ways in which these nets hook up with perception and action would differ. For instance, the "dog net" would be active when throwing sticks for Grete, and the "dinosaur net" would be active when strolling through the dinosaur exhibit of a museum. The point is that intentionality requires not only isomorphism but also some kind of *appropriate causal relation* (sometimes called *input-output relations*) to the world. Not only must the representational structures in the mind or machine be isomorphic to what they represent, but since, as our two examples show, any mental structure will be isomorphic to a host of different things (and vice versa), a representational structure, in order to represent some state of affairs, must typically be triggered by that state of affairs, or something like it, and must typically trigger behavior appropriate to that state of affairs, or something like it.

These two features of an information-processing system may not be sufficient for it to be intentional. Many philosophers argue that more is necessary as well (Searle

1980). But they seem clearly necessary, and they seem at least to be central features of the representational power of human and artificial information-processing systems. Reflection on the nonsufficiency of isomorphism for representation raises the possibility, however, that isomorphism might not even be necessary for intentionality. Some cognitive scientists reflect on the capacity of connectionist networks to employ distributed representations. They note that there is sometimes no obvious feature of these representations that is isomorphic to what is represented, and they would argue that intentionality requires only reliable causal relations of particular kinds between representational states and perception, action, and other representational states, and that no further isomorphism requirement must be attached.

One final characteristic of information-processing systems remains to be discussed: their *modularity on functional dimensions*. This is a point about what kinds of parts these systems decompose into. Compare three objects: an anvil, an automobile engine, and a story-understanding computer. In order to understand the "behavior" of the anvil under various stresses or weather conditions, we can simply decompose it into a set of adjacent regions and investigate their behavior. But suppose that we tried to explain how an automobile engine works by dividing it exhaustively into a set of adjacent one-inch-cubes and then explaining (1) the behavior of each cube and (2) their interactions. Imagine what some of these pieces would contain. One might include part of a piston, part of an intake valve, some empty space, and a bit of the wall of a cylinder. Another might include a piece of carburetor, a piece of air filter, and part of a wing nut. Many cubes would be largely empty. There would be no way to explain the operation of the engine taking this set of cubes as its fundamental parts.

It would be much better to decompose the engine into its "natural" modules: the fuel system, the electrical system, the ignition system, the exhaust system, and so forth, to explain the behavior of each of these systems (perhaps by decomposing them into their natural components), and then to characterize the interactions among these systems. The interesting thing about this strategy (the only one capable of providing an explanation) is that the components into which it divides the engine will in general be related not spatially but functionally. They subserve the same or related functions, and these functions are hierarchically arranged: the fuel system delivers fuel to the cylinders; the carburetor (a component of the fuel system) mixes the gasoline and air; the needle valve assembly controls the amount of gasoline admitted to the carburetor; and so forth. This kind of explanation of how something works is called a *systematic explanation* (Haugeland 1978).

This kind of functional organization is characteristic of information-processing systems as well. The only difference is that in information-processing systems, unlike automobile engines, the parts of the system, their functions, and the ways in which they are interconnected are characterized intentionally, that is, by reference to their representational properties. For instance, a chess-playing computer would decompose not into adjacent one-inch cubes, but into such things as a position decoder, a move generator, a look-ahead device, a tree-pruning routine, position-evaluation routines, and so forth. Each of these components is characterized functionally, rather than physically. Indeed—and this is an important feature of information-processing systems—there is a sense in which it does not matter what physical stuff the components are made of as long as they generate the right output for each input. Moreover, each function, and hence each device, is characterized intentionally. This, then, is the essence of an information-processing system: a system of representations and representation

manipulators that decomposes functionally, and whose functions and components are characterized intentionally. Cognitive science is the attempt to understand the mind as just such a system.

Later in this chapter we will turn to some of the specific philosophical problems that arise from thinking of the mind as an information-processing system (henceforth IPS). Much of the philosophy of cognitive science is taken up with those problems. But first we consider the general structural features of IPSs more carefully to see what it means to think about thinking from an IPS point of view.

The Structure of Cognitive Science

Once we adopt an IPS view of the mind, we think of cognitive processes (deciding, planning movement, retrieving a memory), cognitive states (believing that cognitive science is fun, desiring a cold drink), constructing a visual percept in response to light impinging on our retinas, or solving a difficult puzzle like Rubik's cube as manifestations of a complex set of computational operations on neurally encoded symbols, carried out by a complex IPS, of whose operations we are largely unaware. Those symbols represent not only the things about which we are consciously thinking but also a host of items used internally to the system, of whose very existence we are unaware, such as texture gradients or stack heights.

On this model, our cognitive states—our beliefs, desires, moods, hopes, and fears—are states of this IPS. Exactly what this means is a matter of some dispute (see section 8.3). But the rough idea is that just as a computer's moving its pawn to king's four is, when carefully examined, just an informational characterization of a particular physical state of that computer (voltage high on such and such a line, and so forth), your moving a pawn to king's four is just a way of informationally characterizing your physical state, including perhaps the movement of your arm, as well as the current pattern of neural firings. Cognitive psychology attempts to elucidate the nature of the information processes that mediate between our neural wetware on the one hand, and our beliefs and other conscious states on the other.

This approach invites speculation concerning the medium in which all of these information processes are represented and carried out. The physical symbol system hypothesis, developed in chapters 1 and 2, identifies information processing with symbol manipulation, which involves structured representations and structure-sensitive operations. Fodor (1975, 1987) has extended this view to argue that there is an internal language of thought, innately specified, by means of which all humans represent the world to themselves. Other researchers, inspired by connectionist modeling, argue that human information processing might not involve, properly speaking, computations over symbols at all, despite the fact that it can *mimic* such processing (Churchland 1989; Smolensky 1988). Given its centrality to thought about the nature of mental representation and mental processes, this debate over the internal medium of thought—whether there is one, and if so what it is—has become increasingly salient in the foundations of cognitive science.

Understanding the IPS model of the mind makes it clear just why AI plays such a central role in cognitive science. After all, in the absence of such a model, it would be strange to lump neuroscience, psychology, linguistics, and the philosophy of mind, all seemingly about humans, with AI, a branch of computer science, seemingly about machines. But if the mind is understood as an IPS, as an abstractly characterized formal structure for manipulating representations, then it would seem that it (or portions of it)

can in principle be implemented on a digital computer. Consequently, by studying particular programs running on machines, AI can be seen as a domain for experimenting with cognitive models of the mind in order to divine the structure of human programs. This enterprise would be incoherent in the absence of the IPS model of the mind but is perfectly natural within that model.

Given the IPS viewpoint, the cognitive science discipline that might not seem to fit (if one adopts a classical, as opposed to a connectionist, model of thought) is neuroscience. Although at times we have said that in a sense it doesn't matter what hardware (or wetware) a program runs on, from the IPS viewpoint there is a sense in which it can make a good deal of difference. A chess-playing program on a supercomputer might also run on a personal computer or on a set of filing cards manipulated by thousands of clerks. It might generate identical output for identical input and decompose identically in each of these implementations. But it will run at very different speeds on these devices, taking months to generate each move when implemented by clerks and cards, minutes on a personal computer, and milliseconds on a supercomputer. It might be still faster on a dedicated chess machine, a computer designed solely to play chess. Understanding just how, in detail, the program is implemented by these various systems, and what accounts for their performance characteristics, would be an interesting task. This is one motivation for cognitive neuroscience—to find out how our "software" is implemented on our wetware, and how this implementation affects our cognitive performance. Furthermore, the design of a machine determines that certain programs will be more efficient than others on that machine. As an example, consider a machine with a fast adder and a slow multiplier (and suppose that these functions were represented directly on bits of hardware). Then it might turn out that for certain problems, it would be faster to compute a product by means of a series of addition operations than by a single multiplication operation. Given data about how fast an unknown program ran on such a machine, and a knowledge of its hardware, we might be able to get some important clues about the structure of the program. Similarly, if neuroscientists can tell us interesting things about the strengths and limitations of our nervous systems, from an information-processing standpoint, this, together with performance data, might yield valuable clues about the architecture of the programs we "run." Finally, of course, if neuroscience can provide a radically different model of a computing device (for example, a connectionist model), we might be led to rethink the very model of computation that undergirds the computational model of mind. Some cognitive scientists (for instance, P. M. Churchland 1984, 1989; P. S. Churchland 1986) urge just such a neuroscience-based approach to the study of mind. Remembering that there are currently many viable approaches both to cognitive science and to the philosophy of mind, some computational and some not, it is now time to investigate the philosophical problems raised by conceiving of the mind as an information-processing device of some sort, and by conceiving of a research program in this way.

8.3 Ontological Issues

Ontology

The *ontology* of mind is the study of the nature of psychological states and processes and their relation to physical states and processes. We will consider four central ontological problems raised by cognitive science. These are by no means the only

interesting ontological problems posed by the field, but they are among the most far-reaching and intriguing. First, we will consider what has become cognitive science's version of the mind-body problem: the problem of specifying the kind of relation that holds between psychological and physical events in representational information-processing systems such as human beings and artificially intelligent computers. Second, we will consider how to interpret information-processing theories of human intelligence. Are psychological processes really carried out by the brain in some kind of biological analog of LISP code, or are the programs written by psychologists and computer scientists engaged in cognitive simulation merely useful calculational devices for predicting our behavior? Third, we will ask specific questions about the nature of certain kinds of contentful psychological states such as beliefs, desires, hopes, and fears. Fourth, we will ask what kind of account cognitive science should offer us of the felt quality of our inner experience.

Functionalism

It is useful to begin a discussion of the relation of psychological to computational to physical states with a discussion of *Turing machines*. A Turing machine is a simple kind of computing machine. It is usually described as comprising a *tape* (of any length), divided into discrete cells, upon each of which a single character is written (usually a 0 or a 1); a *read/write mechanism* capable of reading the character on a given cell of the tape, writing a new character, and moving one cell in either direction; a *finite list of internal "states"* the machine can be in; and a *machine table* prescribing for each possible machine state, and each character that the machine might scan while in that state, what character it should write on the tape, which direction it should move after writing that character, and what state it should shift into. These components of the Turing machine are depicted schematically in figure 8.4. In this figure the model Turing machine is in state 2, scanning a 1, and so will print (that is, in this case, leave alone) a 1, move left one cell, and go into (that is, remain in) state 2.

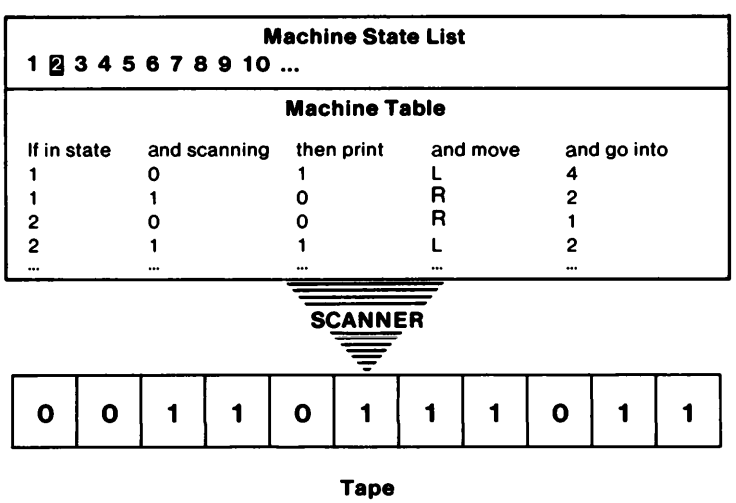

Figure 8.4
Diagrammatic representation of a Turing machine

Despite its simplicity, in terms of the tasks it can accomplish, the Turing machine is the most powerful computational device possible. It is almost certainly true that any computational process that can be performed can be performed by a Turing machine, and it is certainly true that any computation that can be performed by a digital computer can be performed by a Turing machine. In fact, because of the existence of a so-called *universal Turing machine*, a machine that takes a coded version of other Turing machines as inputs, and then emulates their behavior, it is further true that this one machine, the universal machine, itself possesses all of the computational power that any computing machine can possess.

How are Turing machines relevant to an account of the relation between the mind's information-processing states and the brain's biological states? The Turing machine has given cognitive science a persuasive model of what this relation might be and of what the nature of mental states might be. The generic term for the theories inspired by this model, which are embodied to some extent or other by cognitive science, is *functionalism*. In the following three sections we will distinguish several different varieties of functionalism and ask which, if any, constitute plausible theories of the nature of mind. The varieties of functionalism we will consider are *machine functionalism*, *psychofunctionalism*, and what we will call *generic functionalism*.

Machine Functionalism The simplest Turing machine model of the relation of psychological to biological states is that adopted by *machine functionalism*. The machine functionalist notes that the Turing machine is both a physical system and an abstract computing device. Whenever it is in a particular physical state, it is also in a particular machine state, and it is performing some particular calculation (say, adding two numbers or emulating some other Turing machine). And there is no great mystery, no "mind-body problem," about how the physical machine manages to be at the same time a computing machine—that is, about how its machine or computational states are related to its physical states. If someone were to ask how this merely physical machine could possibly be performing the "mental operation" of adding, we would simply point out that each machine state of the system just is a physical state of the system under a computational description.

Machine functionalism asserts that the same might be true of human beings. After all, a Turing machine can represent any information-processing system humans instantiate. They, like us, are finite (though unbounded) physical systems. Hence, if we are physically instantiated information-processing systems, as cognitive science would have it, then we are functionally equivalent to some Turing machine. Since for a Turing machine to be in a particular machine state (its analog of a psychological state) is for it to be in a particular functionally interpreted physical state, it is overwhelmingly plausible to assert that for us, to be in a particular psychological state is to be in a particular functionally interpreted physical state (presumably a biological state of the central nervous system). After all, this line of reasoning continues, a Turing machine can be realized in any kind of physical medium, including both metal and neural matter. It would seem that in the case of humans, then, neural states are to be thought of first as machine states and then as psychological states. On this view, the task of cognitive science is to uncover the machine table that characterizes the machines that human beings instantiate. (For the classic exposition of this view, see Putnam 1960.)

Machine functionalism captures the idea that what is essential to the psychological nature of a mental process or state is not its particular physical realization (though this

may be important for various theoretical and practical reasons) but its computational role in the information-processing system. Hence, it provides an account of how people, intelligent Martians, and suitably programmed digital computers could have the same psychological states and processes, simply by virtue of instantiating, albeit in vastly different physical media, the same Turing machine table.

Psychofunctionalism Even this liberal view of the link between the physical and the psychological nature of information-processing states may be too restrictive, however. One essential component of a Turing machine is its fixed finite list of machine states. But even though the number of machine states for any Turing machine is finite, the number of possible computations that a machine can perform is, in general, infinite. Consider the following example. Suppose that we have a Turing machine capable only of performing addition. Call it A. Even though A may have a very simple machine table, involving only a few states, if we choose to characterize A, not by reference to its machine states, but by reference to what we might call instead its *computational states*, we will see that A is in fact capable of being in an infinite number of states. For A might be adding 2 and 3, or it might be adding 666,666 and 994. It might be "carrying 2," or it might be "writing the answer." Although each of these computational states is equivalent to some sequence of machine states and tape sequences, none is identifiable with any single state, and by virtue of the unboundedness of the set of possible sequences of machine states, there are infinitely many possible computational states of A.

The view that human psychological states are to humans as a Turing machine's computational states are to the Turing machine is known as *psychofunctionalism*. Two general types of reasons motivate psychofunctionalism as opposed to machine functionalism. The first concerns the apparent unboundedness of the class of human psychological states we would like our theories to cope with; the second concerns the criteria we would adduce for ascribing psychological states to humans or machines. Let us consider these in turn.

Suppose that human beings are Turing machines. Then human beings have finitely many possible machine states. How many beliefs could you, as one human being, possibly have? Not at one time, of course, or even actually in one lifetime; but how large is the list of possible beliefs that you are capable of holding, by virtue of your psychological makeup? Could you, for instance, believe that 2 is the successor of 1, that 3 is the successor of 2, and so on, for all natural numbers? If you think that for any belief of this form, you could form that belief and hold it, then you think that you could hold infinitely many possible beliefs. And we haven't even gotten past the most elementary arithmetic! Nor have we touched on desires, hopes, fears, and the multitude of subconscious processes necessary to a complete psychology. Considerations such as these motivate the view that the number of possible psychological states that a complete cognitive science must account for vastly outstrips the number of machine states possible for any Turing machine (though any Turing machine of even small complexity is capable of infinitely many computational states). Of course, this does not imply that the brain might be capable of assuming infinitely many physical states. Quite the contrary. The point is that just as a much richer description of a finite Turing machine results from talking about its computational states than from talking about its machine states (in the sense that there are many more of the former than the latter), so a much richer (and hence possibly psychologically more fertile) description of human brains

would result from talking about their computational states than from talking about their machine states.

Even if by some chance there are only finitely many psychological states in which humans are capable of being, there would be reasons to prefer the psychofunctional account of psychological states. Suppose we have two Turing machines, T^1 and T^2. Both are adding machines, and both are made of the same material. But suppose that their machine tables and sets of machine states are different. Whereas T^1 adds by successively incrementing the first addend by 1 the number of times specified by the second addend, T^2 adds by incrementing the second addend by 1 the number of times specified by the first addend. Moreover, they accomplish their tasks using a different set of machine states. Now, it seems fair to say that when these two machines are adding a pair of numbers, they are, in an important sense, doing the same thing. If we count them as "believing" anything about the sums they compute, then they "agree" in all of their "beliefs," even though they share no machine states. Hence, it seems that the computational level of description is, for some purposes, at least, a useful one for the description of Turing machines. Is this true for people?

Here the case seems, if anything, clearer. For even if it seems far-fetched to attribute beliefs to machines, or odd to think that the computational level would be a particularly interesting level of description for them, it is certainly true that among the psychologically interesting facts about people are the things we believe, fear, doubt, and so on. And even if we are, underneath our surface psychology, Turing machines, many different Turing machines could realize these surface states. Further, it seems that our criteria for attributing these states to ourselves and each other have nothing whatever to do with our views about the machine tables underlying our cognitive processes, but instead have something to do with the relations that these states have to other such states, to the inputs we receive from our environments, and to the behavior we produce in response to them. For instance, if you are disposed to say, "Goats are wonderful," to argue vigorously with those who doubt the virtues of goats, to act in an admiring and friendly way toward goats, to infer from the fact that goats are nearby that something wonderful is nearby, and so on, then others will feel pretty comfortable in ascribing to you the belief that goats are wonderful. And this belief, though it may be supported by some set of your machine states, tape states, and so forth, need not be identified with any particular machine state. Moreover, it certainly need not be the case that there is some machine state that is such that anyone who shares your belief is in that state. Psychofunctionalists argue that psychological states bear a relation to those holding them that is analogous to the relation that computational states bear to their Turing machines, and this has seemed a much more liberal, and indeed more plausible, way to apply the Turing machine metaphor to the task of understanding the nature of mental states.

Generic Functionalism As liberal as psychofunctionalism appears in its account of the psychophysical relation, it is possible to develop an account that is still less restrictive, yet still recognizably functionalist. In order to understand this *generic functionalism*, it is necessary to step back from the Turing machine metaphor and to consider what it is that makes an account functionalist in the first place. Both machine functionalism and psychofunctionalism develop the general idea of functionalism using the Turing machine as the leading idea. But that is really not an essential feature of the functionalist approach. The kernel of the approach is really in the insight that psychological

terms such as *belief, desire, pain, memory,* and *perception* need not be understood as some kind of shorthand either for neurophysiological descriptions or for behavioral descriptions.

The claim that these psychological terms stand for neurophysiological descriptions (a view called *central state identity theory,* or *CSIT*) implies that when someone says that John remembers eating vanilla ice cream and that Bill does, too, then that person is claiming that by virtue of sharing a particular memory, John and Bill share a particular brain state. It would, of course, be an unfortunate consequence of this view that only beings who are neurophysiologically like human beings can have psychological states, thereby ruling out in a single a priori stroke the possibility of intelligent (or even sentient) Martians or computers. In some cases it rules out the possibility of two people sharing any states as well. Suppose, for example, that John lost the left half of his brain in an auto accident at an early age, and Bill lost the right half of his in the same accident; fortunately, however, both were young enough that they have completely recovered with no loss of function, so that despite their apparently similar attitude toward vanilla ice cream, they in fact share no psychological properties.

Thinking that our psychological vocabulary is a set of shorthands for behavioral descriptions is, of course, behaviorism. On this view, to ascribe to both Mary and Sue the desire for a pet unicorn would be to assert that they share a set of behavioral dispositions, including the tendency to chase and attempt to capture any passing unicorns, to say things like "I wish I had a pet unicorn," to search the "Pet Store" section of the Yellow Pages tirelessly for a store carrying unicorns, and so forth. Suppose, though, that Mary is so painfully unassertive that she never openly expresses what she wants, and moreover that she has never heard of the Yellow Pages; and suppose that Sue is a brash and cosmopolitan individual who makes all her wishes known. On the behaviorist account, Mary and Sue could not possibly share this desire (or probably any other, for that matter). Clearly, the behavioral account is inadequate.

The Scylla of CSIT has this much in its favor: it lets Mary and Sue share a desire, despite their personality differences, and gives a fairly straightforward answer to any question about the nature of the psychophysical relation. Unfortunately, it delivers the wrong answer on John and Bill, and it rules out intelligent Martians and computers altogether. The Charybdis of behaviorism has this in its favor: it lets Bill and John share an attitude, despite their different neurophysiological makeup, by virtue of being disposed to say and do the same relevant things. But it fails where Mary and Sue are concerned (and elsewhere, as we have seen).

Functionalism navigates the narrow strait by giving each side its due. The functionalist agrees with the behaviorist that the connections between psychological states and the organism's input and output are central to that state's nature, and that psychological states are independent of particular physical realizations in particular organisms or machines. But the functionalist also agrees with the identity theorist that it is important to look at the inside of the organism, and the interrelations of internal states, in assigning psychological predicates to physical correlates. The functionalist differs from both in suggesting that the right way to understand the mind-body relation is via what is called the *token-identity* theory, that is, that particular (or *token*) psychological events are to be identified with token physical events. Both the behaviorist and the identity theorist subscribe to the stronger *type-identity* theory, which holds that each kind (or *type*) of psychological event is to be identified with a type of

physical (for the identity theorist) or behavioral (for the behaviorist) event. And it is the greater flexibility of the token-identity position that provides the compass that guides the passage.

We have digressed at this length because it is important to get a picture of the general position staked out by functionalism in order to see that although the models of machine functionalism and psychofunctionalism inspired by Turing machines might be strong versions of this view, they are not the only possible versions. Generic functionalists can deny that particular psychological states are to be identified with particular computational or machine states of some Turing machine used to model the mind, while allowing that token psychological states are to be identified with token physical states under one or another scheme for identification, without being committed to the particular scheme. Generic functionalists will agree with their brand-name cousins that whatever this scheme is, it must pay attention to the functional role that the psychological state in question plays in the system being studied—that is, to its relations to inputs, outputs, and other internal states. In short, generic functionalists accept all of the general tenets of functionalism while withholding commitment to any particular version of the mapping from the psychological to the physical. This has the advantage of "looking before leaping" in the absence of concrete evidence for the stronger claims of the brand-name versions and therefore appeals to those of a conservative temperament; but it has the disadvantage of not suggesting a particular research program or line of investigation for determining just how to map psychological states onto physical states and therefore is unappealing to those who want a bold conjecture to test.

Whether generic functionalism is compatible with radical connectionism is a matter of some debate. In a network with hidden units and distributed representations, a psychological state, say, believing that the queen of England is the richest woman in the world, would be stored somehow in the network's set of connection weights. Different learning histories that included the acquisition of this belief would produce different weights. In fact, it is quite possible that two learning histories would produce different values for every single weight in the network. Nevertheless, it is conceivable that some functional property of the weight vector might be discovered that is a necessary and sufficient condition for the possession of the belief. For example, the property might concern the propensity of the network to produce certain pattern of activation over its hidden units when the belief is relevant to the current situation. It is possible, then, that one could be a generic functionalist and a connectionist at the same time (though neither of the stronger versions of functionalism we have considered would be available to a connectionist). On the other hand, either we might fail to discover any such property, or we might find that all the properties that seem to be correlated with the belief are too distributed and temporally unstable to qualify as the belief itself. Generic functionalism would seem to required a belief to be a stable, internal, computationally discrete property, capable of playing a causal role in cognition (Ramsey, Stich, and Garon 1991). Finally, a property of the weights that we wish to identify with the belief might sustain that interpretation only in the context of appropriate input-output relations and by itself deserve no special interpretation at all (to anticipate the considerations in favor of naturalism to be discussed below). Such results might lead to the conclusion that functionalism of any type and connectionism are incompatible and represent genuine alternative to each other. Another alternative, introduced in chapter 2, is to argue that any plausible connectionist model of cognition

will turn out to be an implementation of a classical model (Fodor and Pylyshyn 1988). This debate may be with us for some time.

Interpreting the Theories We have, then, a sketch of the ways in which one can think of the relation between the mind and body if one adopts the computational view of mind embodied by cognitive science. But there are different ways to adopt scientific theories. On the one hand, one could adopt a cognitive theory as literally true and assert that the processes and structures it posits are actually "in the head" (a *realistic* interpretation). On the other hand, one could just suppose that the theories make true predictions about such things as behavior, reaction times, error patterns, and so forth, but not assert that the structures they posit actually reflect psychological reality (an *instrumentalist* interpretation). Before we explore these positions in any detail, let us digress somewhat. Suppose that we, as cognitive scientists, are confronted with a chess-playing computer, playing tolerably good chess. Our task is to offer a theory explaining how the machine works, a theory that will enable us not only to understand its play but also to predict as well as possible what its next move will be in particular situations.

Briefly, considering distinctions drawn earlier, there are three general strategies we could adopt (described in more detail in Dennett 1971). We could describe transistor by transistor, wire by wire, and pixel by pixel, how the current flows through the machine, how it is affected by the depressing of keys, and how it results in the change in luminosity of various regions of the video display (the *hardware strategy*). Or we could abandon that electro-Herculean task in favor of describing line by line, subroutine by subroutine, the program the machine runs in order to play chess, explaining how it encodes board positions, how it represents the values of various parameters, what mathematical operations are performed on these parameters, and so forth (the *program strategy*). Or, dismayed by that prospect as well (quite a daunting one even for a professional programmer), we might offer a theory something like this: "The machine evaluates the current position, looking for threats to pieces, possible forks, and discovered checks. It takes care of those first. Then it looks ahead about two moves, evaluating each possible position according to the balance of material and threats to pieces. It especially worries about losing its major pieces. And one more thing: It knows lots of openings, but only about five moves of each" (the *mentalistic strategy*).

What are the relative advantages and disadvantages of these strategies? First, if (and that's a big "if") the hardware strategy could be made to work, it would give an accuracy of predictive power unmatched by either of the other strategies. The hardware strategy could even predict (something the other strategies could not even do in principle) when smoke would come out of the back of the machine. The problem, of course, is that such an explanation is impossible in practice; moreover, although it tells how this machine really works in one sense, it fails utterly to tell how the machine plays chess. To see this, let us consider another machine, running exactly the same program, but made out of wooden gears instead of silicon chips. Any explanation of how one machine plays chess should be an explanation of how the other plays as well, since they run exactly the same program. But the hardware explanation of our silicon-based computer will be irrelevant to—and certainly false of—its wooden cousin; hence, whatever it does explain, it does not explain our machine's chess-playing ability per se.

The program strategy avoids this problem since it would assign the same explanation to the two physically dissimilar but computationally equivalent computers. Moreover, the explanation would be easier to come up with. These are this strategy's principal advantages. On the other hand, it has a few disadvantages of its own. First, though it is infinitely easier to come up with a program explanation of the abilities of such a system than it is to come up with the corresponding hardware explanation, it is still very difficult, and the explanation might be so complicated that it would provide no real insight into the ability at all. Second, there is much that the program explanation will be unable to handle, such as machine malfunctions that are perfectly amenable to a hardware explanation but are simply outside the scope of the program strategy. Third, the program approach encounters a problem analogous to the problem of rigidity that the hardware approach encounters, though it emerges at a slightly higher level. Consider two machines, alike in hardware, that run programs that implement the same general chess-playing strategies but are written in different programming languages, using different types of underlying subroutines, data structures, and control. Ideally, the explanation of how they play chess should be the same for both. But that would require a "higher," more abstract level of description than the program strategy.

That more abstract level, of course, is what we have called the mentalistic strategy. Here, instead of talking about transistors or cogs, subroutines or addresses, we talk about plans, goals, desires, beliefs, knowledge, and so forth. The disadvantages of this strategy compared with the first two are plain: its predictions will be far less exact, and its explanations in particular cases that much more suspect. It will be completely unable to handle both hardware malfunction and software "bugs." But its advantages are impressive as well. A reasonable amount of close observation will yield reasonably good theories at this level, and the theories will be relatively easy to test and to implement for prediction. Most of all, explanations at this mentalistic level will generalize to all machines using the same kinds of strategies.

Let us focus on the program and mentalistic explanatory strategies, since these are the explanatory strategies most characteristic of cognitive science. A cognitive scientist interested in how people play chess would be primarily interested in understanding what kinds of strategies they employ, or, if the scientist was operating at a more fundamental level, what kind of "program" they are running. Now, the relevant question is this: When interpreting a theory expressed at one of these levels, do we interpret claims that particular programs are being executed, or that particular goals and intentions are being acted on, as possibly literally true, and the processes they posit as in some sense real (a realistic interpretation)? Or do we interpret these theories only as useful predictive instruments, with no claim to real truth, but only to usefulness in prediction and explanation, perhaps pending the development of an explanation of the relevant phenomena in terms of processes with a better claim to reality, perhaps a hardware theory (an instrumentalistic interpretation)?

Both positions can be defended. On the one hand, the advantages of the mentalistic strategy argue in favor of a realistic interpretation of cognitive theories. These theories capture important functional commonalities between human beings, perhaps between human beings and intelligent computers. If there are features that humans share with other physically distinct information-processing systems that explain important aspects of the functioning of both, this argues that such features ought to be treated as real properties of persons qua intelligent organisms. According to this line of

reasoning, theories that explain our behavior by virtue of these processes ought to have every claim to truth.

On the other hand, the instrumentalist would counter, one could treat all of the arguments in favor of adopting the mentalistic strategy, or indeed the program strategy, as supporting the usefulness of these perspectives for predicting the behavior of complex systems but still assert that they say nothing whatever about the truth of program or mentalistic theories or about the reality of the processes the explanations posit. Indeed, the instrumentalist might continue, the fact that, as we ascend the hierarchy of abstraction from the hardware to the mentalistic approach, we lose considerable accuracy and scope of predictive power is strong evidence that what we are doing is trading truth for convenience, accepting a good instrument for human purposes instead of a clumsy, though literally accurate, one. Therefore, the instrumentalist concludes, when we develop program and mentalistic theories of the mind, or of artificially intelligent systems, what we are doing is developing increasingly sophisticated instruments whose accuracy can only be vouchsafed by realistically interpreted theories at the hardware level.

Though some have adopted this instrumentalistic attitude, most cognitive scientists and philosophers of cognitive science accept some version of a realistic interpretation of the theories of cognitive science. They grant that neural matter and silicon are real and that they are the substrata of the higher-level phenomena that cognitive science is interested in explaining and describing in its cognitive theories. But, they argue, all of this does not impugn the reality of the higher-level phenomena supported by hard- or wetware. Those phenomena, too, are real, for they can be shared by objects of radically different hardware constitution and are therefore in some very important sense independent of hardware phenomena. Cognitive scientists' theories study these structures and are true by virtue of making true claims about these abstract structures and processes.

Even though the instrumentalist-realist dispute is very much open in cognitive science (Dennett 1978, 1982; Stich 1983), in what follows we will assume a realistic interpretation of cognitive theories and hence that whatever psychological or computational information processes or states cognitive science requires actually exist, if cognitive science is to be viable.

We have done a good deal of ontological spadework: we have developed the outlines of the functionalist view of the mind that underlies the cognitive science approach, and we have seen what it would be to pursue cognitive science with a realistic interpretation of its theories. We are now in a position to examine particular kinds of psychological phenomena—to ask just how cognitive science should understand them and how they fit into a functionalist philosophy of mind, and what constraints a coherent philosophical account of these phenomena might place on the shape of cognitive theory.

We will take two broad classes of psychological states as examples for the remainder of this ontological investigation. First, we will examine the psychological states that philosophers call *propositional attitudes*. These are states such as belief, desire, hope, and fear that seem to have as their contents *propositions*, or assertions about the world. When, for instance, you *believe that snow is white*, the clause *that snow is white* is a proposition and appears to be the content of your belief. Belief is but one attitude you might take toward that proposition. You might also *doubt*, *fear*, or *hope* it. Propositional attitudes are interesting to cognitive science because they are *relational*. They seem to

involve a certain relation of the individual to the world, or at least to a proposition about the world. Second, we will examine *qualia*, or the felt character of psychological states, such as *what chocolate tastes like* or *what red looks like*. These states are interesting to cognitive science for the opposite reason: they seem to be *nonrelational*. Examining the ontological problems posed by these two classes of phenomena should give us a good feel for the range of ontological issues raised by a realistic, functionalistic theory of mind and hence for the information-processing approach to the study of mind generally. We begin with the propositional attitudes.

Propositional Attitudes

Cognitive science is concerned with propositional attitudes because cognitive science sees its task as explaining the cognitive processes and states of people and other intelligent information-processing systems. Human beings (and perhaps some other intelligent information-processing systems) at least appear to have propositional attitudes. Cognitive science hence owes us, if not an explanation of these states, then at least an explanation of why we appear to have them. And even if the propositional attitudes of which we are immediately aware, such as belief, desire, and the rest, were explained away, instead of merely explained, the problem they pose would remain. Part and parcel of the idea of an information-processing system is the idea that the states of such a system, physical though they may be, are interpretable as having content, even if they are only representing something as mundane as arithmetic operations. On this account, then, even a calculator might be representing the fact that $3 + 4 = 7$, and we would still have to explain just what it is about the machine that constitutes its states' having the content that they do.

Nevertheless, it is not at all clear just what, if anything, cognitive science ought to be expected to tell us about propositional attitudes. For one thing, it is not yet clear whether propositional attitudes are the kinds of things that will appear in a recognizable form in a complete psychology (in the way that gold does both in commonsense talk and in chemistry) or whether they are merely creatures of the commonsense world that have no place in mature science (as the class of precious metals per se has no place in chemistry). Nor is it clear what the boundaries of the domain of cognitive science will be. For all we know now, a cognitive theory may eventually be developed that is adequate to explain a wide range of phenomena, but belief may fall naturally into the domain of a noncognitive portion of psychology. So positions on the place of propositional attitudes in cognitive science vary widely: some hold that cognitive science must provide a complete information-processing account of what belief is and of how it is related to such things as memory and behavior; some that certain aspects of belief will be amenable to cognitive explanation but that others might require a more sociological treatment; some that belief itself is a commonsense notion that will have to be replaced with a more refined concept for the purposes of science; and some that belief is at bottom an incoherent notion that has no place in an accurate description, cognitive or otherwise, of the psychological world. In the discussion that follows we will consider a range of possible approaches to providing cognitive explanations of the nature and role of the propositional attitudes, recognizing that these approaches do not exhaust the possibilities and that it is not obvious to what extent belief is a proper subject of cognitive theory. Such philosophical theorizing about the propositional attitudes may, however, contribute significantly to characterizing the limits and nature of the enterprise of cognitive science by delimiting the nature and range of phenomena

to which it is suited and by sketching the form that a theory would have to take in order to accommodate at least these states and processes.

Propositional attitudes (henceforth simply "beliefs") are problematic for a number of reasons. For our purposes, the central problem is that they look like the kinds of things that should be identified and grouped according to their content, but real scientific and commonsense difficulties stand in the way of doing that.

An example will begin to spell out these difficulties. Suppose that Kilroy is the world's leading cognitive scientist, a renowned goatherd, and Sam's next-door neighbor. Betty knows nothing of Kilroy's glorious caprine successes but has long admired his work in cognitive science. Sam, on the other hand, though blissfully ignorant of Kilroy's career in cognitive science, is in awe of his champion goats. Now, suppose that as Sam and Betty are enjoying a beer at the local dive, in strolls Kilroy. Betty comes to believe *that the world's greatest cognitive scientist has entered the bar.* Sam, on the other hand, comes to believe *that the world's greatest goatherd has entered the bar.* They are, of course, both right, and what makes each right is that each correctly believes *that Kilroy has entered the bar.* The very same fact makes both of their beliefs true. Both of their beliefs are about Kilroy. In fact, a mind-reading bartender would truthfully report that both believe *that Kilroy has entered the bar.* And there's the problem. On the one hand, if the content of their belief is what matters, they do seem to believe the same thing. On the other hand, if the functional role that belief plays in their internal information-processing system is what matters, they do not believe the same thing, since although Betty's belief is connected with other beliefs about cognitive science, Sam's is connected with other beliefs about goats. Hence, they have different relations to other beliefs, inputs, outputs, and so forth, with the result that on any functionalist account they are different beliefs. Therefore, if cognitive science is to realize the dream of a functionalist account of the mind as an information-processing system, and if it is to account for beliefs in this scheme—to treat beliefs, not as classified by their content, but as classified by their functional role in the internal economy of the information-processing system—the question then is, How can cognitive science do this, while at the same time doing justice to the obvious fact that beliefs are beliefs just because they are about things?

The approach to belief that attempts to identify beliefs with particular internal information-processing states is called *individualism.* Two general types of individualism can be distinguished in recent work on the problem of belief: *methodological solipsism* (Fodor 1980; Putnam 1975a; Stich 1983) and *naturalistic individualism.*

Methodological Solipsism Solipsism as a metaphysical thesis is the position that nothing exists outside of the mind of the solipsist. It is hence a very lonely doctrine, and not one frequently defended anymore. Understood methodologically, rather than metaphysically, however, solipsism is more plausible and has more adherents (most of whom recognize each other's existence). The idea is this: we can study mental states and processes without paying any attention at all to the external world they ostensibly represent, and indeed without even assuming that it exists. Rather than explicitly denying the existence of an external world, which would be both absurd and beside the point, we can, the methodological solipsist asserts, ignore the external world for the purposes of cognitive science, in particular, for the purposes of characterizing and attributing beliefs to subjects. The way to do this, the methodological solipsist continues, is to restrict our cognitive science to discussing formal (that is, computationally

characterized) operations on formal (that is, characterized only by reference to their shape) tokens, or states of the organism or computer, and explicitly refusing to discuss any possible interpretation or meaning those states or processes might have for the system in its environment.

Methodological solipsism is motivated by the observations that an information-processing system has no access to the world except through its beliefs and that its processing certainly cannot be sensitive to the interpretations assigned to its internal states. If it is operating on a symbol—say, the English word *you*—the processor will do whatever it does to it regardless of who happens to be in front of it—in fact, regardless of whether anybody happens to be in front of it. Information-processing systems process information by manipulating what are to them meaningless symbols, according to physically determined rules. What makes what they do information processing is that we can later interpret those symbols, states, and processes as meaningful, in light of what we know about the organism's interaction with its environment.

In developing the view motivated by these intuitions, we begin by characterizing the states and processes of our information-processing system purely formally, as a set of uninterpreted formal symbols manipulated by the system according to a set of uninterpreted formal rules, like a complex game or a highly systematic, but possibly meaningless, computer program written in a very abstract computer language. This forms the core of our cognitive theory of the information-processing system in question. Each state of that system, each belief, doubt, perception, intention, and memory, will eventually be identified with one of these as yet uninterpreted states. Then we interpret. We try to assign meanings to the smallest symbols and processes of the system in such a way that the entire system, under this translation scheme, turns out to be functioning sensibly, having mostly true beliefs, making mostly good inferences, and interacting plausibly with its environment. When we succeed in this interpretive task, we are done.

The feature of this approach that deserves emphasis is this: when Betty believes that the world's greatest cognitive scientist has just entered the bar, what is happening as far as cognitive science is concerned is that Betty has processed (in the way that is interpreted as belief in a good cognitive theory of Betty's processor) a formal string of mental symbols (ultimately neurally represented) that are interpreted (in the light of the total behavior of Betty's information-processing system) as a representation of the fact *that the world's greatest cognitive scientist has just entered the bar.* The same can be said for Sam when he believes *that the world's greatest goatherd has just entered the bar.* Neither of them has a representation that is in any essential way connected with Kilroy. Their respective internal cognitive states would presumably be shared by two quite different people, who upon seeing Bill Clinton enter the bar, and having bizarre beliefs about him, believe that the world's greatest cognitive scientist, and the world's greatest goatherd, have entered the bar. On this account, the representations have only to do with the concepts involved, and they mean what they do regardless of what they refer to in the world, if anything. So Kilroy himself plays no part in the cognitive story about Sam's and Betty's beliefs. He does, however, happen to explain the truth of both of their beliefs, by virtue of his happening to satisfy both descriptions that are the best interpretations of the descriptive terms in Sam's and Betty's respective belief-symbols. Thus, methodological solipsism is able to explain how cognitive science could account for the representational character of cognitive states while referring only to internal information processes in characterizing the nature of those states.

A further argument for methodological solipsism is due to Stich (1983), who calls it the argument from the *autonomy principle*. The autonomy principle states that the proper matter for cognitive explanation includes only those states and processes that are realized entirely within the physical bounds of the organism. For example, someone's *being to the west of the World Trade Center* (ignoring the fact that this is not even plausibly a cognitive property) would be ruled out by the autonomy principle as the kind of property with which cognitive science should concern itself. Stich uses what he calls the *replacement argument* to defend the autonomy principle:

> Suppose that someone were to succeed in building an exact physical replica of me—a living human body whose current internal physical states at a given moment were identical to mine at that moment. And suppose further that while fast asleep I am kidnapped and replaced by the replica. It would appear that if the crime were properly concealed, no one (apart from the kidnappers and myself) would be the wiser. For the replica, being an exact physical copy, would behave just as I would in all circumstances. Even the replica himself would not suspect that he was an imposter. But now, the argument continues, since psychology is the science which aspires to explain behavior, any states or processes or properties which are not shared by Stich and his identically behaving replica must surely be irrelevant to psychology. (Stich 1983, 165–166)

This argument needs some refinement to handle certain properties, such as those determined by social relations, but these need not concern us now. The point of the example should be clear. Stich's replica's psychology must be the same as his, and so it must be that the only physical properties that make a difference to Stich's psychology are properties of his body. If that is true, then there is no need for a cognitive theory to pay attention to anything outside Stich's body, and that is what methodological solipsism comes to—that for the purposes of cognitive science, an organism's information-processing states can be characterized without reference to their meaning or their connection with the external world.

This view of mental states as essentially uninterpreted information-processing states that can be identified and explained by cognitive science without reference to their content, but that derive content as a result of our interpreting them in light of the way they and the organism or machine to whom they belong are embedded in the world, has gained much favor within cognitive science, particularly among linguists, computer scientists, and philosophers. But some (for example, Pylyshyn 1980; Bach 1982; McGinn 1982, 1990) are uneasy and suggest that more attention needs to be paid to the meaning of internal states than methodological solipsism permits. They agree with the methodological solipsist that beliefs are internal information-processing states of the individual but deny that they can be identified and explained without looking beyond the individual. The methodology they propose is hence a kind of *naturalism*, by virtue of paying attention to the organism's situation in and relation to nature, but is *individualistic*, in that it continues to view the states themselves as located firmly within the bounds of the individual. We will now examine considerations that lead some to adopt naturalistic individualism.

Naturalistic Individualism Pylyshyn (1980) has noted that certain kinds of explanation might be difficult, if not impossible, to provide in a methodologically solipsistic cognitive science. Suppose, for instance, that we ask the solipsist to explain Mary's behavior:

> It simply will not do as an explanation of, say, why Mary came running out of a certain smoke-filled building, to say that there was a certain sequence of expressions computed in her mind according to certain expression-transforming rules. However true that might be, it fails on a number of counts to provide an explanation of Mary's behavior. It does not show how or why this behavior is related to very similar behavior she would exhibit as a result of receiving a phone call in which she heard the utterance "this building is on fire!", or as a consequence of hearing the fire alarm, or smelling smoke, or in fact following any event interpretable as generally entailing that the building was on fire. The only way to ... capture the important underlying generalisation ... is to ... [interpret] the expressions in the theory as goals and beliefs....
>
> Of course the computational [methodologically solipsistic] model only contains uninterpreted formal symbols.... The question is whether the cognitive theory which that model instantiates can refrain from giving them an intentional [meaningful] interpretation. In the above example, leaving them as uninterpreted formal symbols simply begs the question of why these particular expressions should arise under what would surely seem (in the absence of interpretation) like a very strange collection of diverse circumstances, as well as the question of why these symbols should lead to building evacuation as opposed to something else.... What is common to all of these situations is that a common interpretation of the events occurs.... But what in the theory corresponds to this common interpretation? Surely one cannot answer by pointing to some formal symbols. *The right answer has to be something like the claim that the symbols represent the belief that the building is on fire....* (Pylyshyn 1980, 161; emphasis added)

The point of Pylyshyn's argument is fairly straightforward. A good cognitive explanation of behavior that is motivated by beliefs ought to explain how those beliefs are related to the behavior and to the circumstances that give rise to them. If the beliefs are characterized by the theory as uninterpreted symbols, and if believing is characterized as an uninterpreted process in the believer, then the theory cannot explain their connection either to behavior or to stimulation—or for that matter, to other beliefs. In any real explanation, this objection goes, the content of the belief plays a role. The symbols in Mary's head cause her behavior *because* they represent the fact that there is fire, and any symbols that did not represent that fact would by themselves not explain her behavior. The conclusion that a naturalistic individualist draws is that in a cognitive theory internal information-processing states must be identified by their content, and in order for this to happen, one must of course examine their connections not only to other cognitive states and processes but also to the organism's environment.

This argument charges methodological solipsism with being a useless research strategy. Another line of argument in favor of naturalistic individualism denies the very coherence of the solipsistic strategy. Methodological solipsism insists that information-processing states and processes are to be taken by cognitive theory as uninterpreted formal states and processes and that they are to be identified without paying attention to any relations between the organism and the environment. Now (the naturalist points out) we all agree, as cognitive scientists, that information-processing states are not to be identified physically, for then we could not generalize about information-processing systems realized in physically different substrata, such as human beings and artificially intelligent but perhaps functionally equivalent computers.

The challenge is then posed to the methodological solipsist: given something you have reason to believe is an information-processing system and whose behavior you wish to explain as a cognitive scientist, your task is to decide which of its physical states and processes are going to count as functional or computational states and processes. And of course you cannot, for the reasons we have just discussed, simply stare at the neurons and figure it out. Well, the naturalist continues, I can think of only one way to do it: watch the organism interact with the environment, see how the neural stuff acts when confronted with particular types of stimulation—when the organism performs certain kinds of actions, and solves certain kinds of problems—and interpret the states accordingly. And that, it will be agreed, is a very naturalistic strategy. (But see Fodor 1987 for a determined, though controversial, argument for the compatibility of such a naturalistic research program with a solipsistic psychology, and Garfield 1991 for a reply.)

The point is that simply to make the initial move from the physical level of description to the functional or computational level, as we must in order to do cognitive science at all, is to interpret the system, and the only way to get the data that justify a particular interpretation is to pay attention to naturalistic phenomena. On this account, then, methodological solipsists are wrong in two ways: first, they are wrong in thinking that there is such a thing as an uninterpreted formal description of a physical system, and second, they are wrong in thinking that solipsistic data alone could justify even a minimal interpretation of the states of a physical system as information-processing states.

We have seen powerful arguments for both versions of individualistic interpretations of the propositional attitudes. Which position, if either, is in fact correct is still a hotly debated issue concerning the foundations of cognitive science. But before leaving the subject of the propositional attitudes altogether, we will take a brief look at nonindividualistic accounts of belief. Such accounts are offered by Burge (1979, 1982), McGinn (1990), and Garfield (1988, 1990, 1991), and considered by Stich (1983).

Nonindividualistic Conceptions Nonindividualists take naturalism one step further. They agree that in order to specify the nature of any belief, it is necessary to talk about its content, and that it is impossible to talk about the content of any belief without paying attention to the world outside, including such things as the causes of the belief, the things the belief is about, and the meanings of the words the speaker uses in formulating the belief. Nonindividualists draw a further moral from this need to pay attention to naturalistic data. They infer that belief itself may be a naturalistic phenomenon, that is, that it may be essentially a relation between an organism or machine and the world, rather than a state of the individual organism or machine itself.

An analogy will help to clarify and motivate this point. Individualists treat beliefs as something like internal sentences. The methodological solipsist differs from the naturalistic individualist only in that the former thinks that we can tell what sentences are in someone's head just by looking inside, whereas the latter thinks that we need to look around at the world as well. But suppose that believing *that roses are red* is not so much like writing "Roses are red" in the "belief register" of a person's brain as it is like being related to what "Roses are red" means. The inscription "Roses are red" by itself is not a sentence about flowers. You can see this by imagining a swirl of gases in a far-off nebula, or billions of hydrogen atoms scattered across light-years of intergalactic space, that happen to have the same shape as an English inscription of "Roses

are red." These things make no assertions about flowers. What makes an inscription a sentence expressing a thought about the world is instead its relation to a language and to a community of users of that language. A sentence's being a statement is hence a relational fact about that sentence, much as Harry's being a brother (where Harry is a person who has a brother) is a relational fact about Harry. It is not a fact about Harry per se, in isolation from his environment; rather, it is a fact about one relation between Harry and his environment, in particular between Harry and his family.

Note, for instance, that nobody could tell by just examining Harry—not even his doctor performing the most thorough physical—that he is a brother. Nor, says the anti-individualist, could anyone tell—even by means of the most thorough neural examination—what someone believes. This is just to say that believing *that roses are red* is a relational fact in the same way that an inscription's meaning is. To hold that belief involves a relation among a believer, the corresponding behavior, roses, redness, and a language. On this account, belief is more like brotherhood than like height. It is not a characteristic of the individual, but one of the relations that individual bears to the world.

Those cognitive scientists who adopt this view of belief take one of two attitudes toward the place of belief in cognitive science. Either they decide that belief is not the right kind of thing for cognitive science to study and that it should concern itself only with individualistic phenomena, or they decide that cognitive science must be broadened to encompass not only the nature of the internal information processing of organisms and machines but also their information-theoretic relations to their environments. Both of these approaches involve certain attractions; both are fraught with difficulty. On the one hand, banishing belief is motivated, if belief turns out to be relational, for the reasons suggested by the autonomy principle. But banishing belief seems to involve banishing a central psychological phenomenon from the domain of psychology, and it is not clear what would be left for cognitive science to study. Broadening the purview of cognitive science to encompass relational facts about information-processing systems seems attractive in that it offers the greatest promise of explaining a wide range of cognitive phenomena. On the other hand, there is much to be said for focusing on the already difficult, but somewhat circumscribed, domain of individual information-processing systems in isolation. It should also be noted that such a naturalistic view of belief provides a very natural way of accommodating connectionist models of cognition that defy functionalist or otherwise symbolic interpretation with a realistic view of the propositional attitudes. This may well be despite the fact that no distributed connectionist state of your brain by itself can be identified as the belief, say, that the queen of England is the richest woman in the world. But even if individualism is false, that would not impugn the reality of your belief that she is. For that belief may well comprise a complex set of relations you bear by virtue of that distributed state to external things, such as sentences of English and Elizabeth II, among others.

One final remark about the place of propositional attitudes in cognitive science before we turn to qualia. We have now seen good arguments not only for identifying propositional attitudes both solipsistically and naturalistically but also for construing them individualistically and nonindividualistically. Indeed, it appears that their very nature as states that connect the organism to the world gives them a Janus-like character. On the one hand, in order to figure in internal information processing, they seem to be necessarily individualistically construed; on the other hand, in order to have

content, they seem to be necessarily relational. Without content they seem to lack explanatory power, and without autonomous internal existence they seem psychologically and computationally inert. Therefore, some have suggested, perhaps the right conclusion from these conflicting considerations is that the concepts of belief, and of propositional attitudes generally, are incoherent—in other words, that despite the folklore about human beings as believers, doubters, hopers, and fearers, we in fact are never in any of those states, simply because there is no such thing as a propositional attitude. On this view, to say that you believe that roses are red is as false as it is to say that you are Santa Claus or that you live in a round square house. Here the central task of cognitive science is to construct the notion of an information-processing system in a way that involves no states such as belief at all. This view is developed in various ways by Stich (1983), Ramsey, Stich, and Garon (1991), P. M. Churchland (1984), and P. S. Churchland (1986). Just how such a view would look, and just how it would account for our persistent belief in belief, is not at all clear. Nor is it clear that analogous problems about content will not be raised for the structures that supplant belief in such a theory. The place of propositional attitudes in the computational theory of mind is far from settled, but it is clear that given the central role that the notion of a contentful state plays in cognitive science, it will be important to resolve these problems.

Qualia

The difficulties raised by the propositional attitudes derive from their relational character. It is intriguing that the other class of states thought to raise special problems for cognitive science, the qualia, are thought to raise special problems precisely because they are not relational in character.

The word *quale* (plural *qualia*) is the philosopher's term for the "felt" or "experienced" character of mental states. For instance, although *tasting chocolate* is not a set of qualia, *what chocolate tastes like* (more precisely, *what it feels like to taste chocolate*) is a set of qualia. In order for chocolate tasting to take place, there must actually be some chocolate in rather close proximity to the taste buds of the taster. We could even imagine a mechanical chocolate taster tasting chocolate while experiencing nothing (having no qualia), perhaps only examining the chocolate for the FDA. On the other hand, we could imagine experiencing what it is like to taste chocolate (having chocolate qualia) in the absence of any chocolate, by means of hypnosis or drugs. So the qualia that normally accompany a particular functionally characterized state are at least conceptually separable from the state itself. Whether they are also in fact separable—that is, whether these states can in fact occur without their corresponding qualia, and whether qualia associated with a state can in fact appear without the corresponding state—is another question, one perhaps both philosophical and empirical.

Some psychological states have no intrinsic qualitative character. Believing that the earth revolves around the sun or doubting that the moon is made of green cheese has no particular qualitative character, though each may have a particular functional character. In general, propositional attitudes seem to have no typical qualitative character, but perceptual states, and perhaps emotions and moods, appear to be typically qualitative. Of course, many questions suggest themselves at this point. Are moods and emotions qualitative because they involve some kind of perceptual states, or do they have qualia all their own? Do all perceptual states have associated qualia, or are there some kinds that do not? But we will not ask these questions. Rather, we will ask how qualitative

states differ from propositional attitudes, what possible problems they pose for cognitive science, and whether cognitive science needs to worry about them at all.

The central difference between qualitative states and propositional attitudes is that attitudes always seem to involve a relation between their subject (the believer, hoper, doubter, and so on) and the object of the attitude (the proposition believed, doubted, hoped to be true, and so on). The relation is not easy to characterize, but it seems nonetheless clear that when someone believes something, that person is related by the belief relation to *something*. The qualitative character of a perceptual state, on the other hand, seems to be a *monadic* property of that state, that is, a simple fact about that state of mind, not involving its relation to anything else. Examples of other monadic properties are *being red, weighing one hundred pounds,* and *being a dog*. These are properties that things have that do not involve their relations to other things, as opposed to properties like *being to the left of, being heavier than,* or *being the favorite animal of,* which essentially involve relations to other things. Some argue that the propositional attitudes are monadic properties of persons (Quine 1960; Sellars 1968) and others argue that they are not (Fodor 1978; Stich 1983), but discussion of this view would take us far from both cognitive science and our concern with qualia.

Why Qualia Are Problematic In order to see the possible problems that qualia raise for cognitive science, we must focus on the functionalist theory of mind that we have seen to underlie cognitive science. What makes functionalism a plausible theory of mind is that it offers a good way to identify the psychological states of a natural or artificial system with its physical states: namely, in terms of the relations that they bear to each other, and to the input and output of the system. And these relations are all that matter. On the functionalist view, the intrinsic properties of the state, such as what kind of material the state is realized in, or how long it lasts, or how much noise the system makes getting into it, are irrelevant to its psychological descr ptidn. All that is important here are the relations that the physical state to be described psychologically bears to other psychologically describable physical states of the system.

Those who have suggested that qualia pose special problems emphasize this essential role that the relational (as opposed to the intrinsic) properties of states play in functionalism. They contrast the plausibility of construing propositional attitudes in this way with what they suggest is the implausibility of construing the apparently nonrelational qualia in this way (Block 1978, 1980a; Block and Fodor 1972). A few of the examples often used to make this point will help:

> Imagine a body externally quite like a human body, say yours, but internally quite different. The neurons from the sensory organs are connected to a bank of lights in a hollow cavity in the head. A set of buttons connects to the motor-output neurons. Inside the cavity resides a group of little men. Each has a very simple task: to implement a "square" of a reasonably adequate machine table that describes you. On one wall is a bulletin board on which is posted a state card, i.e., a card that bears a symbol designating one of the states specified in the machine table. Here is what the little men do: Suppose the posted card has a 'G' on it. This alerts the little men who implement G squares—"G-men" they call themselves. Suppose the light representing input I_{17} goes on. One of the G-men has the following as his sole task: When the card reads 'G' and the I_{17} light goes on, he presses output button O_{191} and changes the state card to 'M'. This G-man is called upon to exercise his task only rarely. In spite of the

> low level of intelligence required of each little man, the system as a whole manages to simulate you because the functional organization they have been trained to realize is yours. A Turing machine can be represented as a finite set of quadruples (or quintuples, if the output is divided into two parts)—current state, current input; next state, next output. Each little man has the task corresponding to a single quadruple. Through the efforts of the little men, the system realizes the same (reasonably adequate) machine table as you do and is thus functionally equivalent to you.
>
> I shall describe a version of the homunculi-headed simulation, which is more clearly nomologically possible. How many homunculi are required? Perhaps a billion are enough; after all, there are only about a billion neurons in the brain.
>
> Suppose we convert the government of China to functionalism, and we convince its officials that it would enormously enhance their international prestige to realize a human mind for an hour. We provide each of the billion people in China (I chose China because it has a billion inhabitants) with a specially designed two-way radio that connects them in the appropriate way to other persons and to the artificial body mentioned in the previous example. We replace the little men with a radio transmitter and receiver connected to the input and output neurons. Instead of a bulletin board, we arrange to have letters displayed on a series of satellites placed so that they can be seen from anywhere in China. Surely such a system is not physically impossible. It could be functionally equivalent to you for a short time, say an hour. (Block 1978, 278–279)

Now, it is argued, though it might be plausible that such a homunculi-headed body, whether its homunculi are internal, as in the first case, or external, as in the second, shares all of your propositional attitudes, including, naturally, the belief that it is not a homunculi-head, it would be *implausible* to think that such a creature would share your qualia. Suppose, for instance, that you are completing the last hundred meters of a marathon, and the Peoples' Republic of China (and its robotic input-output device) is in the process of functionally simulating you. It, like you, believes that it is finishing the race. Its robot, like you, is sprinting, or staggering, toward the finish line. Its cognitive processes, like yours, are, by hypothesis, slightly addled. Granting a functionalist account of belief, all of this seems perfectly plausible (give or take a bit of science fiction). But could the Peoples' Republic of China (or its inanimate, remote-controlled robot) feel the same pain (or elation) that you feel at the end of a grueling race? It might believe that it is in intense pain, but what would it be for one billion people and a robot, no one of whom is individually in pain (in any relevant sense), to collectively feel pain, as a result of the inputs to a robot to which they are only connected via walkie-talkies and satellites?

It is the bizarreness of this suggestion that leads many to suggest that qualia pose a special problem for cognitive science. The problem appears to be that (1) qualia are psychological phenomena and are essential to many psychological states (like being in pain, for instance, or seeing red)—hence that they are within the domain that cognitive science ought to cover—but that (2) unlike the propositional attitudes, which are amenable to the relational account offered by functionalism, the qualia of psychological states are monadic, intrinsic features of those states, and not functional attributes. This is demonstrated, the argument continues, by the fact that there can be functionally identical systems, one of whom has qualia (you) and one of whom does not (Peoples'

Republic of China + robot). This is often called the *absent qualia problem*. Hence, the argument concludes, qualia are not functionally characterizable. Hence, states with an essential functional character are not functional states. Hence, not all human psychological states are functional states. Hence, functionalism is not an adequate account of the mental. Hence, insofar as cognitive science is committed to functionalism (and we have seen that it may be rather deeply committed to it), cognitive science is in trouble.

There are two general strategies that a functionalist cognitive scientist could adopt in replying to this qualia-centered charge: deny that cognitive science ought to concern itself with qualia, or meet the argument head on and show that the purported counterexamples (the homunculi-heads) are not real counterexamples, either because they are impossible or because they in fact have qualia. We will consider each of these replies in turn.

Banishing Qualia The task of cognitive science, when applied to human psychology, is to explain and characterize human psychological phenomena as cognitive. In particular, though all psychological states have some noncognitive properties (such as being realized on a digital computer or in a human brain), these noncognitive properties are not the business of cognitive science. Furthermore, by virtue of being functionalist, cognitive science has an account of what it is for a property to be cognitive: it is for that property to be a functional, information-processing state of the system in question. If functionalists could argue persuasively that qualia not only are not functional states (just what the absent qualia objector urges against them) but also are not really part of the domain of cognitive psychology at all, they would have a strong reply to the qualia objection—in effect, that claim (2) is right—qualia are not functionally characterizable or explicable—but claim (1)—that they are within the purview of cognitive science—is wrong, and hence functionalists should no more be worried about their inability to explain them than they are worried about their inability to explain the common cold.

The argument begins by noting that many psychological states are qualitative. But that no more guarantees that they are necessarily qualitative, when considered just as psychological states, than the fact that some psychological states are neurological guarantees that they are necessarily neurological when considered just as psychological states. Hence, some properties of any psychological state—in particular, physical properties—though they serve as underpinnings for psychologically important properties of that state, are themselves irrelevant to the psychological identity of the state. Furthermore, the very examples that both the functionalist and the absent qualia objector offer suggest that functionally identical states might differ qualitatively, in just the way that we have seen that they can differ physically. Now, functionalism independently provides a good theory of the nature of psychological states. Given this fact about the possible physical dissimilarity between cognitively identical states, these examples should suggest only that qualitative character is to a psychological state just as physical character is to it—a character that many, or perhaps even all, states have, but one that is accidental to their cognitive nature and hence not within the purview of cognitive psychology. On this account, the right way to think about psychological states is as functional states, which typically have qualitative character but whose qualitative character is not essential to their psychological nature. Explaining and characterizing this qualitative character is no business of cognitive science on this account, except insofar as we are concerned to explain the particular physical

realization that a psychological state might have. Such an account might vary from system to system. Of course, one would have to add that explaining the belief that one is having qualia of a given type would fall within the purview of cognitive science on this account, since beliefs are functional states according to the theory we have been exploring. But on this view, beliefs about one's own qualia no more entail the need for an explanation of the qualia than beliefs about elephants call for a theory of elephants. To the extent that one is comfortable exiling qualia from the domain of psychology, this is an attractive position (Churchland and Churchland 1981).

Qualia Functionalized If one wanted to keep qualia within the domain of cognitive science and therefore wanted to defend a functionalist account of qualia, one might argue something like this: The very examples that defenders of qualia employ against cognitive science are incoherent. These examples suppose that there are possibly two kinds of states, both of which are typically caused by the same sorts of things, and which typically cause the same sorts of beliefs that one is in a state of a particular qualitative character. For example, we are to imagine that on finishing the race described earlier, you are in state Q_1, which is caused by running a marathon and which causes one to believe that one is in pain, but the Chinese homunculi-headed robot is in state Q_2, which is also caused by running a marathon and which also causes one to believe that one is in pain. However, state Q_1 produces a true belief that one is having pain qualia, whereas state Q_2 produces a false belief of the same kind. Now, what could account for this difference?

There seem to be two possible ways to answer this question. One could simply say that there is no difference—that any state that has such and such causal properties by virtue of such and such functional relations is qualitative—hence, the homunculi-heads and their cousins the computers, intuitions to the contrary, have qualia. Or one could insist that there is some nonfunctional fact about genuinely qualitative states such as Q_1, not shared by such ersatz states as Q_2, that accounts for their being genuinely qualitative.

There is considerable reason to argue that the homunculi-heads do, appearances and intuitions to the contrary, have qualia. After all, what more could you say about a pain, or a sensation of red, other than that it is the very thing that is typically caused by (...) and gives rise to all the (very complex) set of beliefs, desires, and so on that (...) and dispositions to do (...). Obviously, "..." will become very complicated and will be no easy matter to spell out, but that is the task of cognitive science and is what makes it hard and interesting. And imagine trying to convince someone (something) that has just been injured (broken), and is sincerely telling you that it is in pain, and is acting as though it is in pain, that it is mistaken in thinking that it is in pain, because, though it has all of the right beliefs, desires, and behavior, and for all of the right reasons, it lacks qualia.

On the other hand, if defenders of qualia maintain that there is something nonfunctional about qualitative states that distinguishes them from their ersatz counterparts, they have the difficult task of telling us what that is. One reason that it is difficult is this. The distinguishing quality must be something that makes a difference in people's cognitive lives—that makes real qualitative states different from ersatz qualitative states (Davis 1982)—or else there is no difference between the two, and the argument is over. Presumably, this difference must involve the ability of genuine qualitative states to produce some effects on our beliefs, desires, and so on, that ersatz states

cannot; otherwise, what would the difference come to? But if that is true, the functionalist can reply, then since our beliefs are functional, we can define genuine qualitative states functionally in terms of the beliefs they cause, and ersatz states functionally in terms of the different beliefs they cause. Hence, on this view, if there is a difference between real and ersatz qualitative states, it is a functional difference, and so qualia are functional. Therefore, the apparent examples of nonfunctional differences between real and ersatz qualia (the internal and external homunculi-heads) are not real examples, since it is impossible for a system to be functionally equivalent to but qualitatively different from us (Shoemaker 1981, 1982).

This reply exerts considerable pull unless one adopts the view (explicitly denied by the defender of functionalism in the last paragraph) that there is a difference (hard to capture though it may be) between real and ersatz qualitative states that does not show up in the interactions of these states with functionally characterized states but is instead an irreducible, introspectible, and monadic property of these states themselves. The defender of qualia can argue that just as many things (the sight of an elephant, drugs, fever) can produce in us the belief that we are seeing an elephant, so many things (pain, dreams, ersatz pain) can produce in us the belief that we are in pain. But it would not follow from the fact that ersatz and real pain bore the same relations to all functional states that they are, or feel, the same, just as it would not follow from the fact that some drug and the sight of an elephant both typically cause one to believe that one is seeing an elephant that the drug and the elephant are the same thing. And, the defender of qualia would conclude, as mental states, qualia are within the domain of cognitive science, for to feel pain is not, like being in a particular neural state, merely an accidental feature of a psychological state, it is that psychological state itself; hence, if cognitive science is to say anything about the mind, it must say something about qualia, and if functionalism is incapable of capturing qualia, something must change in cognitive science.

This debate is obviously complicated, and it remains one of the interesting areas for philosophical research in the ontology of cognitive science. It is, of course, possible (Searle 1980) that the problems about propositional attitudes that we have raised and the very different-looking problems about qualia are really two sides of the same underlying problem about the ability of functionalism and the computational paradigm to account for meaningful states generally. Those sympathetic with this outlook have suggested that the problem is to be located in the view, essential to the cognitive paradigm, that physical realization is inessential to psychological properties—that the mental can be abstracted from its physical substrate. Those who take a more biological view of the mental reject this assumption; they argue that the psychological is essentially biological and hence that the problems we have seen about propositional attitudes and qualia are simply problems that arise when one mistakenly treats intelligence and mentality as abstract, information-processing concepts. This is a dispute that goes to the heart of the ontological foundations of cognitive science. We have only touched the surface.

8.4 Epistemological Issues

We now turn to the epistemological issues relevant to cognitive science. *Epistemology* is the branch of philosophy concerned with the nature, structure, and origins of knowledge. Traditionally, the major issues in epistemology with which philosophers have

concerned themselves have been the analysis of the concept of knowledge and the nature of the justification of belief. But cognitive science has forced a reconception of what epistemological issues demand attention, and currently issues about the structure and organization of the representation of knowledge are coming to center stage. We will first discuss the nature of what has come to be called the *knowledge representation problem*—the problem of just how to represent large bodies of knowledge in such a way that they can be mobilized to guide behavior and to understand and produce language. We will then examine a few of the special problems about how to understand the concept of knowledge in the context of cognitive science.

The Knowledge Representation Problem

Artificial intelligence (AI) is so central to cognitive science because it both embodies the computational model of cognition and serves as a test of the enterprise. If we could build intelligent information-processing systems on digital foundations, that would show that digital information-processing systems can be intelligent and would provide a powerful reason for believing that humans can be described in that way; should it prove impossible to build such systems, or should serious principled difficulties arise, that would constitute powerful evidence that cognitive science is headed in the wrong direction. Among the most powerful tests of an AI system is its ability to understand natural language.

Understanding natural language requires a vast array of knowledge, and it is important that the knowledge be arranged in such a way that at a moment's notice the system can draw upon just the right bit to help it understand the text it is reading or hearing. Haugeland (1979) drives this point home with a few apt examples:

(1) I left my raincoat in the bathtub because *it* was still wet. (p. 621)

(2) When Daddy came home, the boys stopped their cowboy game. *They put away their guns and ran out back to the car.* (p. 625)

(3) When the police drove up, the boys called off their robbery attempt. *They put away their guns and ran out back to the car.* (p. 625)

Strictly speaking, the italicized texts are all ambiguous. The *it* in (1) could refer to the bathtub, but that is not the most likely interpretation. The second sentences in (2) and (3) are identical. But they mean very different things, and no speaker of English would pause for a moment over the ambiguity. For anyone with the right kind of common sense and linguistic ability, these texts are unambiguous in context. The question is, How do we (and how should an AI system) represent the knowledge that enables us to understand natural language in cases like this in such a way that we can mobilize it just when we need to in order to effortlessly disambiguate texts like these?

This is no trivial problem. Suppose that the way we understand (1) is by retrieving the fact *that putting a raincoat in a bathtub makes sense if the raincoat is wet, but not if the bathtub is.* This would be a pretty strange piece of knowledge to have floating around in our heads. Think of how many others like it we would need to have if this were really how we worked: *that grizzly bears don't like champagne, that there is no major league baseball on Uranus,* and so on ad infinitum. Probably, then, this is not the right account.

We have explored some of the strategies that the field of AI uses in order to try to solve this knowledge representation problem. Frames and scripts are one approach. But

even frames, with all their flexibility and power, are not clearly adequate to tasks like those posed by this collection of mundane texts. As Haugeland (1979) points out, although both a bathtub and a raincoat frame would contain the information that the respective objects could get wet, neither would plausibly contain just the piece of information we need. It would also appear that production or other rule-based systems would have great difficulty with such texts. These systems are even less flexible than frame systems and more prone to being led down "garden paths" of misinterpretation. This kind of problem may be best viewed as akin to a pattern-matching problem. For instance, some researchers claim that recognizing letters, words, or familiar objects is something that is more easily accomplished by connectionist systems than by classical systems. Solving these problems may not require a great many (or even any) explicitly stored statements or rules, but instead a well-tuned network that reliably maps stimulus patterns into actions. It may be that by virtue of a well-tuned cognitive-neural connectionist network, human beings reliably and with no explicit reliance on propositional knowledge or inference rules simply map the "raincoat" situation onto the action of putting the raincoat into the bathtub, and that is all there is to it. Maybe much of our knowledge is like that. This, of course, would represent a dramatic reconceptualization of what and how we know.

The problem of how we in fact represent the myriad bits of information we obviously represent about the world in a way that allows us to find just what we want when we want it is still unsolved. It is a central problem of cognitive science.

Procedural, Declarative, and Tacit Knowledge

In discussing knowledge representation, we have been talking about knowledge almost exclusively as though it is "knowledge that ..." or (to use the philosopher's term) *declarative knowledge*, that is, knowledge of the truth of declarative sentences (for example, "Putting a raincoat in a bathtub makes sense if the raincoat is wet, but not if the bathtub is"). But it is far from clear that all of our knowledge is of this form.

Traditional epistemology distinguishes between *knowing how* and *knowing that*. Though this distinction is not the same as the one psychologists draw between procedural and declarative knowledge, the two are closely related. Much of our knowledge-that is probably encoded declaratively, since much of it is mobilized in controlled processes. Similarly, the kinds of automated, production-style skills we have are typically demonstrated in situations where "know-how" is the most apt characterization of the knowledge in question. An example of a piece of procedural knowledge that is also know-how is *knowing how to ride a bicycle*. We might also know declaratively that a bicycle has two wheels and that we must balance in order to ride it. But it is a very different thing to know that we have to balance it and to know how to accomplish that feat, as any child with training wheels will testify. Not only can we have some knowledge-that without having the corresponding knowledge-how; we can also know how to do something without knowing that we do it in the way that we do. However, these distinctions do not coincide exactly. We may, for instance, know how to solve a tricky puzzle by virtue of representing declaratively a set of rules for its solution, and it may be correct to say that a baby knows that crying will lead to feeding, even if all that is represented is a production rule mediating a highly automated procedure that fires when the baby's stomach is empty. In what follows we will oversimplify somewhat and refer to knowledge-that as declarative knowledge and knowledge-how as procedural knowledge, examining only the cases where the distinction collapses.

Given that some knowledge appears to be declarative and some appears to be procedural, we can begin to ask some interesting questions. Is there some knowledge that is *necessarily* procedural or *necessarily* declarative? Is some knowledge more efficiently represented procedurally or declaratively? Is linguistic knowledge more accurately characterized procedurally or declaratively? Does it make any real difference how we choose to represent a particular item or kind of knowledge? Could all knowledge be represented one way or the other? Is knowing what the word *hammer* means more like knowing that it has six letters or like knowing how to use a hammer? Does the distinction between classical and connectionist models of the architecture of cognitive systems correspond to or crosscut the procedural/declarative or the how/that distinction?

In another version of the argument we have called Ryle's regress, Ryle (1949) argued that procedural knowledge is more fundamental than declarative knowledge—that is, that all declarative knowledge presupposes some procedural knowledge, but not vice versa. In particular, many tasks requiring intelligence, such as reading, problem solving, speaking one's native language, and carrying on a conversation, are guided by procedural rather than declarative knowledge. Ryle was concerned to argue against what he called "the intellectualist legend" according to which in order to do anything intelligently was to do it guided by some internally represented declarative knowledge about the task. Ryle argued that this view was committed to an infinite regress of such data structures. For if to do anything intelligently is to do it in a way guided by some declarative knowledge, then to use the relevant declarative representations for a particular intelligent task intelligently would require using declarative knowledge about which information to use, how to use it, and so forth, and to use that knowledge intelligently would require a further data structure, and so on ad infinitum. To use any information without consulting the relevant meta-information would be to use it unintelligently; hence, the entire operation would be guided unintelligently and therefore would be unintelligent. Hence, Ryle argued, any view of intelligent action that requires that action to be guided by declarative knowledge in order to be intelligent must be misguided.

Of course, the point of this argument is not that there is no declarative knowledge, or even that intelligent behavior is not often guided by declarative knowledge. Plainly, there is and it is. Rather, the point is that it cannot be declarative "all the way down." At some point (perhaps often, appearances to the contrary, at the very top) the regress of declarative representations must bottom out with at least the knowledge of how to use the relevant declarative representations. Since this knowledge cannot be declarative, on pain of the regress, all declarative knowledge presupposes at least the knowledge of how to access and use that knowledge, whereas procedural knowledge presupposes no declarative knowledge. Hence, the argument concludes, procedural knowledge is the most fundamental kind of knowledge.

Not surprisingly, given the role of declarative representations in contemporary approaches to cognitive science, not all cognitive scientists are persuaded by this argument. Fodor (1981) expresses one reply in this way:

> Someone may know how to X and not know how to answer such questions as "How does one X?" But the intellectualist [cognitive] account of X-ing says that, whenever you X, the little man in your head [control routine of the program you run] has access to and employs a manual on X-ing; and surely, whatever is

his is yours. So again, how are intellectualist theories to be squared with the distinction between knowing how and knowing that?

The problem can be put in the following way. Intellectualists want to argue that cases of X-ing involve employing rules the explication of which is tantamount to a specification of how to do X. However, they want to deny that anyone who employs such rules *ipso facto* knows the answer to the question "How does one X?"

What, then *are* we to say is the epistemic relation [way of representing] an agent necessarily bears to rules he regularly employs in the integration of behavior? There is a classical intellectualist suggestion: if an agent regularly employs rules in the integration of behavior, then if the agent is unable to report these rules, then it is necessarily true that the agent has *tacit* knowledge of them. (Fodor 1981, 73–74; reprint of Fodor 1968)

The regress, this line of argument suggests, is generated only if we worry in the wrong way about how the behavior is executed. If we insist, with Ryle, that for the behavior to be executed intelligently is one thing, requiring guidance by declaratively represented rules, and for it to be merely executed is another, requiring no such guidance by representations, then we will fall prey to the regress. But suppose instead that to execute the behavior requires that the system represent something like an internal manual (Fodor's term for a structure of declaratively represented information). Suppose further that the relevant declarative representation is *tacit* (hence unconscious or unavailable to introspection) and that the system is wired so as to access that information in order to execute the behavior. We can then treat the intelligence of behavior as a description of the quality of the information used by the system, or of the procedures that make use of it, and no regress arises.

This reply to the regress involves two key insights. The first is the insight that there can be declarative knowledge that is not conscious, that the organism is unable to articulate—that is, tacit knowledge. To attempt to argue that the knowledge that guides a particular performance is procedural simply on the grounds that the person or machine performing the procedure cannot explain how it does it is to ignore the possibility that much of our knowledge, whether procedural or declarative, is inaccessible to our introspection.

The second key insight behind this cognitive reply to Ryle's regress is that in order for behavior to be guided by an internal declarative representation, it does not follow that the system needs a further declarative representation to guide its access to the first representational structure. At some point, this reply points out, the operation *Get the information necessary for dialing the telephone* can be "wired into" the system, and an executive whose only job is to access the right information at the right times does not have to do much more than recognize those times. Hence, even for behavior to be guided intelligently (that is, by a good representational structure), all that is needed is that a rather dumb executive routine recognize that it is time to activate that particular structure. Then the knowledge contained in that structure can be used by dumb processes to guide intelligent behavior. And that is the way that large computer operating systems work.

This line of argument is certainly plausible, and it may ultimately be the correct reply to the regress argument. However, it is important to note that several key issues are swept under the rug in adopting this reply, issues that are central to the

epistemology of cognitive science. First, if the bulk of the knowledge that guides intelligent behavior is to be represented explicitly, albeit tacitly, in internal "manuals" or some other kind of declarative data structure, with a relatively dumb executive acting as librarian for the system, it is necessary to specify what the content of these manuals will look like. It is well and good to focus on things like dialing telephones, tying shoes, stacking blocks, and other such simple, self-contained operations. But is there to be a manual for conducting a conversation about the weather, for doing literary criticism, or for selecting a movie? What gets put in what manual? This, of course, is a problem that goes beyond merely declarative data structures and can be raised as well for highly procedural knowledge representation systems, such as production systems. It is the problem of what knowledge to put where, of how to organize it for quick access, and of how to design an executive adequate for the access task.

Second, even if there were a way to partition the knowledge we represent into a tidy library of manuals for guiding behavior, it is by no means clear that the "dumb executive librarian" that would have the job of selecting the right manual at the right time would have such an easy job that it could be very dumb. It takes a certain amount of judgment to know whether it is appropriate to take the *Run for Your Life* manual off the shelf rather than, say, the *Fundamentals of Self-Defense* manual. Maybe both are part of the *How to Cope with Danger* manual. Once the books get *that* fat, however, the library loses much of its point, for vast amounts of procedural knowledge will be needed even to help find the right chapter.

Considerations such as these suggest that it is probably necessary when thinking about knowledge representation to think about employing a healthy mix of procedural and declarative strategies for representing necessary knowledge (not only the knowledge necessary for text understanding but also the knowledge necessary for guiding action) and that simply solving the problem of how to represent knowledge about some small portion of the world (or a "microworld") may leave untouched the larger and more fundamental problem of how to represent and organize the large amounts of information necessary to get around in a world that resists tidy compartmentalization. It is also important, these considerations suggest, to remember that to "know" something, whether it be knowledge-how or knowledge-that, is not necessarily to know consciously; it is only to somehow represent the relevant information in a way that makes it accessible for information processing.

Linguistic Knowledge

As an example of the usefulness of the concept of tacit knowledge in cognitive theorizing, let us consider the representation of specifically linguistic knowledge. Two epistemologically interesting claims are often made about linguistic, especially syntactic, knowledge: (1) that although we have no conscious access to them, and although we frequently violate them, we *know* (the as yet scientifically unknown) rules of the syntax of our native language and (2) that our knowledge of certain universal principles and parameters of syntax is innate.

These claims initially sound puzzling since, although the notion of tacit knowledge makes sense, we might think that the only grounds for asserting that a set of rules is represented tacitly in a system is that the system always obeys them. After all, we are by hypothesis denied the evidence of the system's reciting the rules to us. But we do not always obey the rules of our grammar. Given that any knowledge we have of these rules must be tacit, and that the only grounds for attributing tacit knowledge of

a set of rules to a system is that it obeys them, and that we do not always obey the rules of our grammar, why can we nonetheless legitimately say that we tacitly know those rules?

The answer to this puzzle involves distinguishing, as in chapter 6, between our linguistic *competence* and our linguistic *performance*. Our linguistic performance is what we actually do. It depends on many factors that have nothing to do with our linguistic knowledge or with cognitive science: what we know, how tired we are, how much we've had to drink, what music we've been listening to, and so on. Our linguistic competence describes what we are able to do, under ideal conditions, simply by virtue of our knowledge of our language. A model of English competence is a model of the idealized speaker of English. The business of linguistic theory is to explain ideal behavior, linguistic competence, and not the countless deviations from competence occasioned by the slings and arrows of outrageous fortune. (The task of explaining such deviations is left to other branches of cognitive science, such as cognitive psychology, psycholinguistics, and, in extreme cases, neuroscience.) Now, if the task is to explain our linguistic competence, and if the best explanation of our linguistic competence involves suggesting that we follow a highly articulated set of rules, then it seems that we are forced to the conclusion that we somehow represent those rules in a way that enables us to guide our behavior. And that, given the fact that we are completely unable to say what those rules are, suggests that we tacitly know them. Now, of course, this is not to say whether these rules are represented declaratively or procedurally, classically or connectionistically (and whether, if the latter, in a way that actually lets us isolate the representation of particular rules at all), or at what level of analysis their representation is to be found. That is no concern of linguistics, or even of epistemology, but rather of empirical psycholinguistics. What turns out to be the most efficient form in which to represent linguistic knowledge will depend a great deal on other as yet undiscovered facts about the structure of the human information-processing system. But it does seem clear that the facts that we do not accurately follow our grammar and that we cannot articulate it in no way impugn the assertion that we tacitly know it.

This completes our brief survey of some of the major epistemological issues that confront a philosopher of cognitive science. There is clearly much scope for work to be done, but it should also be clear that philosophy has a substantial contribution to make, both to the process of coming to a synoptic understanding of the nature and commitments of cognitive science and to the assessment, reformulation, and revision of cognitive theory and research.

8.5 The State of Cognitive Science

We have surveyed the structure and ontology of cognitive science and some of the epistemological problems it poses. We have developed the idea of an information-processing system and explored the value of that idea as a framework for understanding the mind. It is an extremely fruitful idea. This is obvious from the pace and results of research in cognitive science. It is also remarkable evidence in favor of this approach to the study of the mind that it has sparked such a thorough and exciting convergence of ideas and research among psychology, philosophy, neuroscience, computer science, and linguistics. The view of the mind that emerges is both scientifically and philosophically compelling.

We have also seen that in cognitive science philosophy is not a mere "commentator" on the activities of the other disciplines. Philosophy functions as a team player, helping to define problems, criticize models, and suggest lines of inquiry.

But we have also encountered some outstanding philosophical problems confronting cognitive science. There is the matter of what brand of functionalism looks best as an account of the mind-body relation. Each version has certain advantages, but each is beset with deep philosophical difficulties as well. There is the problem of whether to adopt a realistic interpretation of cognitive theory, and of what the consequences would be. A sound account is needed of the nature both of the propositional attitudes and of qualia. Disquiet about the nature of these complementary classes of psychological states leads to deeper worries about the connection between mind and its physical substrate that penetrate to the very foundations of the cognitive approach. These are not mere conceptual playthings. They are ontological problems that must be soluble if the cognitive approach is coherent. This situation, of course, is not unique to cognitive science. All sciences pose philosophical problems, and the existence of difficulties does not necessarily indicate that those difficulties are insuperable.

The epistemological issues confronting cognitive science also raise a myriad of outstanding issues: the knowledge representation problem, which is both formidable and central to the enterprise; the procedural-declarative issue and the many problems of detail it raises; and fortunately other problems that philosophy has already helped to solve as well as to pose. Again, however, noticing difficulties is not tantamount to noticing certain failure, and nothing we have said about either ontological or epistemological issues could be interpreted at this stage as evidence of the imminent demise of cognitive science, only of much work, and much philosophical work at that, to be done.

Even if we had very good reason to believe that one or more of the problems raised in this chapter was indeed insoluble, and that because of its intractability the information-processing approach to understanding intelligence and human behavior was ultimately doomed, this would not be a reason to give up on cognitive science. After all, Newtonian physics ultimately turned out to be false, but had it not been pursued, relativistic physics could never have been born. Similar analogies can be found in all of the sciences. It is a fact of scientific progress that the advent of each new, more-close-to-true theory or approach is made possible only through the work of earlier scientists pursuing an ultimately false or ultimately unworkable approach. The point of science, including cognitive science, is always to pursue the best research program going, and to push it as far as it will go. It will either turn out to be correct, or, if not, it will almost certainly lead to the discovery of a better approach.

Brainstorms (Dennett 1978) offers a number of essays on topics in the philosophical foundations of cognitive science, focusing primarily on questions concerning intentionality and the interpretation of intentional psychological theories but discussing a number of other related topics as well. For more discussion of the relation of mind to brain and of the role of neuroscience in cognitive science, see *Matter and Consciousness* (P. M. Churchland 1984) or *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science* (P. M. Churchland 1989). For a more detailed view of the relevant neuroscience in a philosophical context, see *Neurophilosophy* (P. S. Churchland 1986); for a more popular treatment, see *Minds, Brains, and Science* (Searle 1984).

Further discussions of philosophical problems raised by AI can be found in "Semantic Engines" (the introduction to Haugeland 1981) and *Artificial Intelligence* (Haugeland 1985), in *Gödel, Escher, Bach: An Eternal Golden Braid*, a wide-ranging, more popular, and often highly entertaining and intriguing treatment (Hofstadter 1979), and in two particularly skeptical treatments of AI, *What Computers Can't Do: A Critique of Artificial Reason* (Dreyfus 1979) and *Minds, Brains, and Science* (Searle 1984). Good discussions of the nature of psychological explanation and the structure of cognitive theory are to be found in *The Nature of Psychological Explanation* (Cummins 1983) and *The Science of the Mind* (Flanagan 1984), and an especially careful treatment of the role of computational models in psychological theory in *Computation and Cognition* (Pylyshyn 1984). For interesting challenges to the reality of such commonsense psychological states as propositional attitudes, see (from the perspective of neuroscience) *Matter and Consciousness* (P. M. Churchland 1984) and *Neurophilosophy* (P. S. Churchland 1986) and (from a more computational perspective) *From Folk Psychology to Cognitive Science: The Case against Belief* (Stich 1983). *Representations: Philosophical Essays on the Foundations of Cognitive Science* (Fodor 1981) and *The Modularity of Mind* (Fodor 1983) offer articulate expositions and defenses of the functionalist, computational model of mind, as does Computation and Cognition (Pylyshyn 1984). *Belief in Psychology: A Study in the Ontology of Mind* (Garfield 1988) offers a critical survey of a number of proposals regarding the attitudes and a defense of a naturalistic account of mind. *Mental Contents* (McGinn 1990) provides a detailed examination of the scope and limits of such a naturalism. For treatments of epistemological issues in cognitive science, see *Cognition and Epistemology* (Goldman 1986) and *Language, Thought, and Other Biological Categories* (Millikan 1985). *Simple Minds: Mental Representation from the Ground Up* (Lloyd 1989) offers a compelling integrated vision of the embodiment of mind in simple organisms and machines, as well as in humans, and a sensitive exploration of the tension between classical and connectionist models.

References

Bach, K. (1982). *De re* belief and methodological solipsism. In Woodfield 1982.

Block, N. (1978). Troubles with functionalism. In Savage 1978. Also in Block 1980b.

Block, N. (1980a). Are absent qualia impossible? *Philosophical Review* 89, 257–274.

Block, N. ed., (1980b). *Readings in philosophy of psychology*. Vol. 2. Cambridge, Mass.: Harvard University Press.

Block, N., ed. (1980c). *Imagery*. Cambridge, Mass.: MIT Press.

Block, N., and J. A. Fodor (1972). What psychological states are not. *Philosophical Review* 81, 159–181. Also in Fodor 1981.

Burge, T. (1979). Individualism and the mental. In French, Uehling, and Wettstein 1979.

Burge, T. (1982). Other bodies. In Woodfield 1982.

Churchland, P. M. (1984). *Matter and consciousness*. Cambridge, Mass.: MIT Press.

Churchland, P. M. (1989). *A neurocomputational perspective: The nature of mind and the structure of science*. Cambridge, Mass.: MIT Press.

Churchland, P. M., and P. S. Churchland (1981). Functionalism, qualia, and intentionality. *Philosophical Topics* 12, 121–145.

Churchland, P. S. (1986). *Neurophilosophy*. Cambridge, Mass.: MIT Press.

Cummins, R. (1983). *The nature of psychological explanation*. Cambridge, Mass.: MIT Press.

Davis, L. (1982). Functionalism and qualia. *Philosophical Studies* 41, 231–249.

Dennett, D. (1971). Intentional systems. *Journal of Philosophy* 63, 87–106. Also in Dennett 1978 and Haugeland 1981.

Dennett, D. (1978). *Brainstorms*. Cambridge, Mass.: MIT Press.

Dennett, D. (1982). Beyond belief. In Woodfield 1982.

Dreyfus, H. (1979). *What computers can't do: A critique of artificial reason.* 2nd ed. New York: Harper and Row.

Flanagan, O. J. (1984). *The science of the mind.* Cambridge, Mass.: MIT Press.

Fodor, J. A. (1968). The appeal to tacit knowledge in psychological explanation. *Journal of Philosophy* 65, 627–640. Also in Fodor 1981.

Fodor, J. A. (1975). *The language of thought.* Cambridge, Mass.: Harvard University Press.

Fodor, J. A. (1978). Propositional attitudes. *Monist* 61, 501–523. Also in Fodor 1981.

Fodor, J. A. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences* 3, 63–73. Also in Fodor 1981 and Haugeland 1981.

Fodor, J. A. (1981). *Representations: Philosophical essays on the foundations of cognitive science.* Cambridge, Mass.: MIT Press.

Fodor, J. A. (1983). *The modularity of mind.* Cambridge, Mass.: MIT Press.

Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind.* Cambridge, Mass.: MIT Press.

Fodor, J. A., and Z. Pylyshyn (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition* 28, 3–71.

French, P. A., T. E. Uehling, and H. K. Wettstein, eds. (1979). *Midwest studies in philosophy.* Vol. 4: *Studies in metaphysics.* Minneapolis, Minn.: University of Minnesota Press.

Garfield, J. (1988). *Belief in psychology: A study in the ontology of mind.* Cambridge, Mass.: MIT Press.

Garfield, J. (1990). *Epoche* and *Sunyata*: Scepticism East and West. *Philosophy East and West* 40, 285–307.

Garfield, J. (1991). Review of Fodor, *Psychosemantics. Philosophy and Phenomenological Research* 52, 235–239.

Goldman, A. (1986). *Cognition and epistemology.* Cambridge, Mass.: Harvard University Press.

Gunderson, K., ed. (1975). *Language, mind, and knowledge.* Minneapolis, Minn.: University of Minnesota Press.

Haugeland, J. (1978). The nature and plausibility of cognitivism. *Behavioral and Brain Sciences* 2, 215–260. Also in Haugeland 1981.

Haugeland, J. (1979). Understanding natural language. *Journal of Philosophy* 76, 619–632.

Haugeland, J., ed. (1981). *Mind design: Philosophy, psychology, artificial intelligence.* Cambridge, Mass.: MIT Press.

Haugeland, J. (1985). *Artificial intelligence: The very idea.* Cambridge, Mass.: MIT Press.

Hofstadter, D. (1979). *Gödel, Escher, Bach: An eternal golden braid.* New York: Basic Books.

Lloyd, D. (1989). *Simple minds: Mental representation from the ground up.* Cambridge, Mass.: MIT Press.

McGinn, C. (1982). The structure of content. In Woodfield 1982.

McGinn, C. (1990). *Mental contents.* Oxford: Blackwell.

Millikan, R. G. (1985). *Language, thought, and other biological categories.* Cambridge, Mass.: MIT Press.

Newell, A., and H. A. Simon (1976). Computer science as empirical enquiry. *Communications of the Association for Computing Machinery* 19, 113–126. Also in Haugeland 1981.

Putnam, H. (1960). Minds and machines. In Putnam 1975b.

Putnam, H. (1975a). The meaning of 'meaning'. In Gunderson 1975. Also in Putnam 1975b.

Putnam, H. (1975b). *Mind, language, and reality: Philosophical papers.* Vol. 2. Cambridge: Cambridge University Press.

Pylyshyn, Z. (1980). Cognitive representation and the process-architecture distinction. *Behavioral and Brain Sciences* 3, 154–169.

Pylyshyn, Z. (1984). *Computation and cognition: Toward a foundation for cognitive science.* Cambridge, Mass.: MIT Press.

Quine, W. V. O. (1960). *Word and object.* Cambridge, Mass.: MIT Press.

Ramsey, W., S. Stich, and J. Garon (1991). Connectionism, eliminativism, and the future of folk psychology. In J. D. Greenwood, ed., *The future of folk psychology: Intentionality and cognitive science.* Cambridge: Cambridge University Press.

Ryle, G. (1949). *The concept of mind.* London: Hutchinson.

Savage, W., ed. (1978). *Perception and cognition: Issues in the foundations of psychology.* Minneapolis, Minn.: University of Minnesota Press.

Searle, J. (1980). "Minds, brains, and programs. *Behavioral and Brain Sciences* 3, 417–457. Also in Haugeland 1981.
Searle, J. (1984). *Minds, brains, and science.* Cambridge, Mass.: Harvard University Press.
Sellars, W. (1968). Some problems about belief. *Synthese* 19.
Shoemaker, S. (1975). Functionalism and qualia. *Philosophical Studies* 27, 291–315.
Shoemaker, S. (1981). Absent qualia are impossible: A reply to Block. *Philosophical Review* 90, 581–599.
Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11, 1–74.
Stich, S. (1983). *From folk psychology to cognitive science: The case against belief.* Cambridge, Mass.: MIT Press.
Woodfield, A., ed. (1982). *Thought and object.* Oxford: Oxford University Press.