CHAPTER 1

Mind

What is a mind?
Could we make a machine with a mind?
What is the relationship between minds and bodies?

1.1 Introduction

In countless movies, computers play a starring role. Some talk in synthesized voices; others write a stream of words on a screen. Some manage spaceships; others, the "brains" of robots, manage their own "bodies." People converse with them, are understood by them, exchange information and greetings with them. Much of this is still science fiction. But real computers advise lawyers on relevant cases, doctors on diagnoses, engineers on the state of atomic reactors. Both the fantasy and the fact would have astonished our grandparents. *Their* grandparents might have thought that this could only be achieved by magic. Yet most of us are getting used to it, taking the silicon age for granted.

Still, a suspicion remains. We human beings have always thought of ourselves as special. We all assume some contrast between the world of material things and the world of spiritual things. If the computer really is a "material mind," then not only must we rethink this distinction, but we have broken it with our own creations. We should be careful to avoid such an important conclusion until we have really thought it through. However natural it seems to take it for granted that computers can think and act, then, we shouldn't just assume it. In philosophy we often find that what we normally take for granted—the "commonsense" point of view—gets in the way of a proper understanding of the issues. So let's see if the way I spoke about computers in the first paragraph is accurate.

I said that they talk. But do they *really* talk in the sense that

people do? It isn't enough to say that they produce something that sounds like speech. Tape recorders do that, but they don't talk. When people talk they mean something by what they say. To mean something, they need to be able to understand sentences. Now I also said that computers understand what we say to them. But do they really? The sounds of our speech are turned into electrical impulses. The impulses pass through the circuits of the machine. And that causes the speech synthesizer to produce sounds. It may be very clever to design a machine that does this, but what evidence do we have that the machine understands?

Well, could a machine understand? There are two obvious responses to this question. The first response I'll call **mentalist**, for the sake of a label. It's the response you make if you think that understanding what people say involves having a mind. The mentalist says:

> Computers can't really understand anything. To understand they would have to have conscious minds. But we made them from silicon chips and we programmed them. We didn't give them conscious minds. So we know they don't have them.

At the other extreme is the response I'll call **behaviorist**. The behaviorist says:

> Naturally, everyone should agree that some computers don't understand. But there's no reason why a computer couldn't be made that does understand. If a machine responds in the same ways to speech as a person who understands speech, then we have just as much reason to say that the machine understands as we have to say that the person does. A machine that behaves in every way as if it understands is indistinguishable from a machine that understands. If it behaved in the right way, that would show that it had a mind.

It is clear why I call this response "behaviorist." For the behaviorist says that to understand is to *behave* as if you understand.

What we have here is a situation that is quite familiar in philosophy. There are two opposing views—mentalist and behaviorist, in this case—each of which seems to have something in its favor, but neither of which looks completely right. Each of these views has a

bit of common sense on its side. The mentalist relies on the common sense claim that machines can't think. The behaviorist relies on the common sense claim that all we know about other people's minds we know from what they do. It looks as though common sense here isn't going to tell us if the mentalist or the behaviorist is right.

In fact, if you hold either of these views you can face difficult intellectual choices. Let's start with a problem you get into if you are a mentalist. Suppose the computer in question is in a robot, which, like androids in science fiction, looks exactly like a person. It's a very smart computer, so that its "body" responds exactly like a particular person: your mother, for example. For that reason I'll call the robot "M." Wouldn't you have as much reason for thinking that M had a mind as you have for thinking that your mother does? You might say, "Not if I know that it's got silicon chips in its head." But did you ever check that your mother has got brain tissue in her head? You didn't, of course, because it wouldn't prove anything if you did. Your belief that your mother has a mind is based on what she says and does. What's in her head may be an interesting question, the behaviorist will say, but it isn't relevant to deciding whether she has thoughts. And if it doesn't matter what is in your mother's head, why should it matter what's in M's?

That's a major problem if you're a mentalist: how to explain why you wouldn't say an android had a mind, even if you had the same evidence that it had a mind as you have that your mother does. Surely it would be absurd to believe your mother has a mind on the basis of what she does and says, yet refuse to believe M has a mind *on the very same evidence*. If it's the evidence of what your mother does that entitles you to believe she has a mind (and not, say, an innate prejudice), then the very same evidence about something else would entitle you to believe that *it* had a mind. This is one line of thought that might lead you to behaviorism.

But if you decide to be a behaviorist, you have problems too. You and I both know, after all, since we both do have minds, what it is like to have a mind. So you and I both know there's a difference between us and a machine that behaves exactly like us but doesn't have any experiences. Unless M has experiences, it hasn't got a mind. The difference between having a mind and operating as if

you've got one seems as clear as the difference between being conscious and being unconscious.

The upshot is this: If you look at the question from the outside, comparing M with other people, behaviorism looks tempting. From the point of view of the evidence you have, M and your mother are the same. Looked at from the inside, however, there is all the difference in the world. You know you have a mind because you have conscious experiences, an "inner life." M may have experiences, for all we know. But if it doesn't, no amount of faking is going to make it true that it has a mind.

We started with a familiar fact: computers are everywhere and they're getting smarter. It looks as though there will soon be intelligent machines, machines that will understand what we say to them. But when we look a little closer, things are not so simple. On the one hand, there is reason to doubt that behaving like a person with a mind and having a mind are the same thing. On the other, once we start asking what and how we know about the minds of other people, it seems that our conviction that people have minds is no better based than the belief that there could be understanding computers. We call someone who asks philosophical questions about what and how we know an **epistemologist**. And if we ask how we know about the minds of other people it seems plain that it is from what they say and do. We simply have no direct way of knowing what—if anything—is going on in other people's minds. But then, if what people say and do is what shows us they have minds, a machine that says and does the same things shows us that it has a mind also. From the epistemologist's point of view, other people's minds and the "minds" of computers are in the same boat.

When we look at the question from the inside, as we have seen, the picture looks different. Someone who looks from the inside we can call a **phenomenologist**. "Phenomenology" is the philosopher's word for reflecting on the nature of our conscious mental life. From the phenomenologist's point of view, M, and all machines, however good they are at behaving like people, may well turn out not to have minds.

From thinking about computers in science fiction we have found our way to the center of the maze of problems that philosophers call the **philosophy of mind** or **philosophical psychology**.

As I said in the introduction, philosophical perplexity is a little like being lost in an old city. It is time now to find our way up that tower to have a look around. We have already been forced back to two of the most fundamental philosophical questions, "What is it to have a mind?" and "How do we know that other people have minds?" So let us put aside the question about M and take up these more fundamental questions directly. At the end of the chapter I'll get back to M, and we'll see then if our trip up the tower has indeed helped us to find our way about.

1.2 Descartes: The beginnings of modern philosophy of mind

The dominant view of the mind for the last three hundred years of Western philosophy has been one that derives from the French philosopher René Descartes, one of the most influential philosophers of all time. His method is to start looking at questions by asking how an individual can acquire knowledge. He starts, that is, by asking how he knows what he knows; and if you want to see the force of his arguments, you will have to start by asking yourself how you know what you know. The fact that Descartes starts with how he knows things marks him as one of the first modern philosophers. For, since Descartes, much of Western philosophy has been based on epistemological considerations.

Descartes' best-known work, the *Discourse on Method*—its full title is actually *Discourse on the Method for Properly Conducting Reason and Searching for Truth in the Sciences*—is written in a clear, attractive style. This may make what he is saying seem simpler and more obvious than it really is, so we need to consider what he says very carefully. Here is a passage from the fourth part of the *Discourse*, published in 1637, where he sets out very clearly his view of the nature of his own self:

> Then, examining attentively what I was, and seeing that I could pretend that I had no body and that there was no world and no place where I was; but that I could not pretend in the same way that I did not exist; and that, to the contrary, just because I was thinking to doubt the truth of other things, it followed quite obviously and quite certainly that I did exist; whereas if I had just ceased to think, although everything else that I had ever imagined had been true, I

> would have had no reason to believe that I existed; I knew from this that I was a substance whose whole essence or nature was only to think, and that had no need for any place to exist and did not depend on any material thing; so that this "I," which is to say my mind, through which I am what I am, is entirely distinct from my body, and even that it is easier to know than my body, and further that even if my body did not exist at all, my mind would not cease to be all that it is.

This passage contains practically every central component of Descartes' philosophy of mind.

First, Descartes is a **dualist**. This means he believes that a mind and a body are two quite distinct sorts of thing, two kinds of what he calls "substance."

Second, what he thinks you really are, your self, is a mind. Since you are your mind, and minds are totally independent of bodies, you could still exist, even without a body.

Third, your mind and your thoughts are the things you know best. For Descartes it is possible, at least in principle, for there to be a mind without a body, unable, however hard it tries, to become aware of anything else, including any other minds. Descartes knew, of course, that the way we do in fact come to know what is happening in other minds is by observing the speech and actions of "other bodies." But for him there were two serious possibilities, each of which would mean that our belief in the existence of other minds was mistaken. One is that these other bodies could be mere figments of our imagination. The other is that, even if bodies and other material things do exist, the evidence we normally think justifies our belief that other bodies are inhabited by minds could have been produced by automata, by mindless machines.

Fourth, the essence of a mind is to have thoughts, and by "thoughts" Descartes means anything that you are aware of in your mind when you are conscious. (The **essence** of a kind of thing, K, is the property—or set of properties—whose possession is a **necessary and sufficient condition** for membership in K. That is, if something has the essential property E, then it belongs to K—so E is sufficient for membership in K; anything that *doesn't* have E doesn't belong to K—so E is necessary for membership.) In other places Descartes says that the essence of a material thing—the property, in other

words, every material thing must have—is that it occupies space. This means that for Descartes the two essential differences between material things and minds are (1) that minds think, whereas matter does not, and (2) that material things take up space, whereas minds do not. Descartes' claim, then, is that what distinguishes the mind from the body is the negative fact that the mind is not in space and the positive fact that the mind thinks.

It is not surprising that Descartes believed that matter does not think. Very few people suppose that stones or tables or atoms have thoughts. But why did he think that minds were not in space? After all, you might think that my mind is where my body is. But if I had no body, as Descartes thought was possible, I would still have a mind. So he couldn't say that a mind *must* be where its body is, simply because it might not have a body at all. Still, if I do have a body, why shouldn't I say that that is where my mind is? If I didn't have a body, that would be the wrong answer; but, as it happens, I do.

I think the main reason for thinking that minds are not in space is that it does really seem strange to ask, "Where are your thoughts?" Even if you answered this question by saying "In my head," it would not be obvious that this was literally true. For if they were in your head, you could find out where they were in your head, and how large a volume of space they occupied. But you cannot say how many inches long a particular thought is, or how many centimeters wide, or whether it is currently north or south of your cerebral cortex.

There is a fifth and final characteristic of this passage that is typical of Descartes' philosophy of mind: throughout the argument Descartes insists on beginning with what can be known for certain, what cannot be doubted. He insists, that is, on beginning with an epistemological point of view.

These are the major features of Descartes' philosophy of mind, and, as I said, this has been the dominant view since his time. So dominant has it been, in fact, that by the mid-twentieth century the central problems of the philosophy of mind were reduced, in effect, to two. The first was a problem M made us think about, **the problem of other minds:** What justifies our belief that other minds exist at all? And the second is **the mind-body problem:** How are we to explain the relations of a mind and its body? The first of these

questions reflects Descartes' epistemological outlook; the second reflects his dualism.

Now, it is just this dualism that raises some of the major difficulties of Descartes' position. For anyone who thinks of mind and body as totally distinct needs to offer an answer to two main questions. First, how do mental events cause physical events? How, for example, do our intentions, which are mental, lead to action, which involves physical movements of our bodies? Second, how do physical events cause mental ones? How, for example, is it possible for physical interaction between our eyes and the light to lead to the sensory experiences of vision, which is mental? And, as we shall see, the answer Descartes gives to these questions seems not to be consistent with his explanation of the essential difference between body and mind.

Descartes' answer to these questions seems clear and simple enough. The human brain, he thought, was a point of interaction between mind and matter. Indeed, Descartes suggested that the pineal gland, in the center of your head, was the channel between the two distinct realms of mind and matter. That was his answer to the mind-body question.

But this theory comes into conflict with Descartes' claim that what distinguishes the mental from the material is that it is not spatial. For if mental happenings cause happenings in the brain, then doesn't that mean that mental events occur in the brain? How can something cause a happening in the brain unless it is another happening in (or near) the brain? Normally, when one event—call it "A"—causes another event—call it "B"—A and B have to be next to each other, or there has to be a chain of events that are next to each other which runs from A to B. The drama in the television studio causes the image on my TV screen miles away. But there is an electromagnetic field that carries the image from the studio to me, a field that is in the space between my TV and the studio. Descartes' view has to be that my thoughts cause changes in my brain and that these changes then lead to my actions. But if the thoughts aren't in or near my brain, and if there's no chain of events between my thoughts and my brain, then this is a very unusual brand of causation.

Descartes wants to say that thoughts aren't anywhere. But, according to him, at least some of the effects of my thoughts are in

my brain, and none of the direct effects of my thoughts are in anybody else's brain. My thoughts regularly lead to my actions and never lead directly to someone else's. We have now reached one central problem for Descartes' position. For it is normal to think that *things are where their effects originate*. (We can call this the **causal account of location**.) And on this view my thoughts are in my brain, which is the origin of my behavior. But if mental events occur in the brain, then, since the brain is in space, at least some mental events are in space also. And then Descartes' way of distinguishing the mental and the material won't work. Let's call this apparent conflict between

a) the fact that mind and matter do seem to interact causally and
b) Descartes' claim that the mind is not in space

Descartes' problem. Once you accept the causal account of location, there are four main ways you might try to deal with this problem.

The first would be to deny that causes and their effects have to be in space. Descartes' is only one of the possible dualist solutions to the mind-body question that takes this approach. Because he thinks that mental and material events interact, even if only in the brain, his view is called **interactionism**. But if you want to keep Descartes' view that the mind is not in space, and if you do not think that causes and effects of events in space have themselves to be in space, you might also try one of the other forms of dualism. There are two kinds of dualism you might try in which the causation goes only one way. You could hold either that mental events have bodily causes but not bodily effects, or that mental events have material effects but no material causes. Each of these positions deserves consideration. But each of these two kinds of dualism claims that minds are both causally active in space and yet somehow not in space themselves. As a result, they need to offer some way of thinking about causation that is very unlike the way we normally think about it.

A second way out of Descartes' problem is to deny that there are any causal connections between mind and matter at all. On this view there are corresponding material and mental realms, which run in

parallel, without any causal interaction. **Psychophysical parallelism**, as this theory is called, certainly escapes Descartes' problem. But we are left with a mystery: why do the mind and the body work together if there is no interaction between them? Psychophysical parallelism says mind and body run in parallel without explaining why.

The third way out of Descartes' problem would be to try a different way of distinguishing mind and matter. If you think that both causes and their effects have to be in space and that mental events have material causes or effects, you cannot maintain Descartes' claim that minds are not spatial. Starting with some new way of distinguishing mind and matter, however, you might still be able to keep dualism, while taking into account the fact that causes have to be in space if their effects are.

But however you distinguish the mental and the material, if you believe they are two different kinds of thing you will have to face the other-minds problem. If your mind and body are utterly distinct kinds of thing, how can I know anything about your mind, since all I can see (or hear or touch) is your body? You brush off the fly, and I judge that you want to get rid of it. But if there is no necessary connection between what your body does and what is going on in your mind, how is this judgment justified? How can I know your body isn't just an automaton, a machine that reacts mechanically, with no intervening mental processes? If you find this thought compelling, you might want to try a solution to Descartes' problem that is not dualist at all.

So the fourth and last way out of Descartes' problem is just to give up the idea that mind and matter really are distinct kinds of thing, and thus to become what philosophers call a "monist." **Monism** is the view that reality consists of only one kind of thing. For monists, beliefs and earthquakes are just things in the world. Things in the world can interact causally with each other, so there's nothing surprising about my belief that there's a table in my way causing me to move the table. The movement of the table is partly caused by the belief. That's no more surprising than a movement of the table caused by an earthquake.

I've suggested that thinking about the other-minds problem might lead you to give up dualism. And if you consider the very evident fact that we *do* know that other people have minds you may be

led, with many twentieth-century philosophers and psychologists to the form of monism called "behaviorism." **Behaviorism**, which we noticed as one possible response to the problem of deciding whether a computer could have a mind, is simply the identification of the mind with certain bodily dispositions. A behaviorist, then, is someone who believes that to have a mind is to be disposed to behave in certain ways in response to input. On one behaviorist view, for example, for English-speakers to believe that something is red is for them to be disposed to say, "It is red," or to reply with a "Yes" if asked the question "Is it red?" And dispositions like this are a familiar part of the world. Being sharp is (roughly) being disposed to cut if pressed against a surface; being fragile is (roughly) being disposed to break if dropped.

There's a strong contrast between behaviorism and Descartes' view. Descartes thought belief was a private matter. That had two consequences. First, that you know for sure what you believe. Second, that *only you* know for sure what you believe. And the trouble with Descartes' view of the mind is that it makes it very hard to see how we can know about other minds at all. For the behaviorist, on the other hand, belief is a disposition to act in response to your environment. If you respond in the way that is appropriate for someone with a certain belief, that's evidence that you have it. Since your response is public—visible and audible—others can find out what you believe. Indeed, as the English philosopher Gilbert Ryle argued in his book *The Concept of Mind*, we sometimes find out what we ourselves believe by noticing our own behavior.

It is a big step from saying that some of our mental states are things that other people can know about, to saying, with the behaviorists, that all of them must be in this way public. Yet one of the most influential philosophical arguments of recent years has just this conclusion. The argument was made by the Austrian-born philosopher Ludwig Wittgenstein, whose work we will discuss again in the chapter on language.

Wittgenstein began by supposing that anyone who believed in the essentially private thoughts of Descartes' philosophy of mind would find it quite acceptable to suppose that someone could name a private experience—one, that is, that nobody else could know about. And indeed, as we shall see in Chapter 3, Thomas Hobbes,

who was an English philosopher who reacted against some of Descartes' ideas, thought that we used words as names of our private thoughts in order to remember them. He called them "marks" of our thoughts. To use marks in this way, someone would have to have a rule that they should use the name just on the occasions where that private experience occurred. Wittgenstein argued that obeying such a rule required more than that there should be both circumstances when it was and circumstances when it wasn't appropriate to use the name. He thought that it also required that it should be possible to *check* whether you were using the name in accordance with the rule. And he offered a very ingenious argument that was supposed to show that such checking was impossible. If Wittgenstein was right, there could be no such "private languages." And his argument is called, for that reason, the **private-language argument**.

1.3 The private-language argument

Wittgenstein's objection to a Hobbesian private language depends, as I have said, on a claim about what is involved in following a rule. His *Philosophical Investigations* begins by introducing the idea of a **language-game**, which is any human activity where there is a systematic rule-governed use of words. One of the conclusions Wittgenstein suggests we should draw from his consideration of language-games is that the notion of following a rule can only apply in cases where it is possible to check whether someone is following it correctly. If someone uses a word or a sentence in a rule-governed way, Wittgenstein argues, it must make sense to ask how we know that they are using the rule correctly; or, as he puts it, there must be a "criterion of correctness."

Suppose, for example, Mary claims to be using the word "tonk" in a language-game. We watch her for a while, and she says the word "tonk" from time to time but we cannot detect any pattern to the way she uses the word. So we ask her what rule she is following. If Mary claims simply to know when it is appropriate to use the word but we cannot discover what it is that makes her use of the word appropriate, then we have no reason to think she is following a rule. Unless we can check on whether it is appropriate for Mary to use the word "tonk," we cannot say that there is a difference between

Mary's following a rule, on the one hand, and Mary's simply uttering a sound at random from time to time, on the other.

Let us now see how Wittgenstein can put the claim that rule following involves a criterion of correctness to use in attacking the Hobbesian private language.

We can start by considering in a little more detail the kind of private use of language that Hobbes thought was possible. Suppose I have an experience that I have never had before. For a **Cartesian** (this is the adjective from "Descartes") there can be no doubt in my mind either that I am having the experience or what the experience is. Still, since it is new, I might want to give it a name, just so that if it ever comes along again, I can remember that I have had it before. So I call the experience a "twinge." I know exactly what a twinge is like, and I just decide to refer to things like that as "twinges." Of course, I cannot show you a twinge and, since I don't know what caused it in me, I don't know how to produce one in you either. My twinge is essentially private: I know about it and nobody else can.

This story seems to make sense. But Wittgenstein thought that if we analyzed the matter a little further, we could see that it does not. Here is the passage where Wittgenstein makes his objection to the sort of Hobbesian private language that I have described.

> Let us imagine the following case. I want to keep a diary about the recurrence of a certain sensation. To this end I associate it with the sign "S" and write this sign in a calendar for every day on which I have the sensation.—I will remark first of all that a definition of the sign cannot be formulated.—But still I can give myself a kind of ostensive definition.—How? Can I point to the sensation? Not in the ordinary sense. But I speak, or write the sign down, and at the same time I concentrate my attention on the sensation—and so, as it were, point to it inwardly.—But what is this ceremony for? for that is all it seems to be! A definition surely serves to establish the meaning of a sign.—Well, that is done precisely by the concentrating of my attention; for in this way I impress on myself the connection between the sign and the sensation.—But "I impress it on myself" can only mean: this process brings it about that I remember the connection right in the future. But in the present case I have no criterion of correctness. One would like to say: whatever is going to seem right to me is right. And that only means that here we can't talk about "right."

Before we try to work out what the argument is that Wittgenstein is making here, we should notice a number of features of the way this passage is written. This passage is rather like a dialogue in a play. Some philosophers, such as Plato, whom we'll discuss in the next chapter, actually wrote philosophical dialogues in order to make their arguments. Wittgenstein doesn't give different names to the people expressing different points of view. Nevertheless you can see that what is going on here is, in effect, a discussion between someone who believes that Hobbes's story makes sense and someone who does not. This means that we have to be careful to decide which of the positions is the one that Wittgenstein is actually defending. In fact, he was defending the point of view of the position which has the last word in this passage: the point of view of the person who says that "this means that here we can't talk about 'right.' " We must try to see what Wittgenstein means by this claim and how he argues for it.

So how does he get to this conclusion? Let's make explicit the fact that two opposed positions are represented here, by identifying each of them with a character. We might as well call one of these characters "Hobbes" and the other "Wittgenstein." Then we can paraphrase this passage as if it were a philosophical dialogue; and, for the sake of concreteness, let's call the sensation a "twinge," as we did before, rather than using Wittgenstein's rather neutral term "S."

HOBBES: For there to be a private language, all that is required is that I associate some word, "twinge," with a sensation and use that word to record the occasions when the sensation occurs.

WITTGENSTEIN: But how can you define the term "twinge"?

HOBBES: I can give a kind of ostensive definition. In an **ostensive definition**, we show what a term means by pointing to the thing it refers to. Thus, suppose we were trying to explain to someone—a person who didn't know English—what "red" meant. We could point to some red things and say "red" as we pointed to them. That would be an ostensive definition of the word "red."

WITTGENSTEIN: But for an ostensive definition to be possible, one must be able to point to something, and in this case pointing is not possible. I cannot point to my own sensations.

HOBBES: Naturally, you cannot literally point to a sensation, but you can direct your attention to it; and if, as you concentrate on the sensation, you say or write the name, then you can impress on yourself the connection between the name, "twinge," and the sensation.

WITTGENSTEIN: What do you mean by saying you "impress the connection on yourself"? All you can mean is that you do something whose consequence is that you remember the connection correctly in future. But what does it mean, in this case, to say that you have remembered it correctly? In order to be able to make sense of saying that you have remembered it correctly, you must have a way of telling whether you have remembered it correctly, a criterion of correctness. And how would you check, in this case, that you had remembered it right?

This is the key step in the argument. Wittgenstein asks Hobbes in effect to consider the question "How do you know, when you say 'Aha, there's another twinge,' that it is the same experience you are having this time?" "Well," Hobbes might answer, "since nothing is more certain than what is going on in your own mind, there can be no doubt that you know."

But if it is possible for you to remember correctly, then it must be possible that you remember incorrectly. After all, according to Hobbes, it is the fact that we may forget an experience that makes names useful as marks. So suppose you have misremembered. Suppose that this experience is in fact not the same experience at all. How could you find out that this was so? And, if you can't find out, what use is the word "twinge"? The name gives you no guarantee that you have remembered correctly, if you have no guarantee that you know what the name refers to.

In order to bring out the force of Wittgenstein's argument, you might argue as follows. Hobbes's idea is that the name can help you remember that you have had the experience before. If it is possible that you have forgotten the experience of the twinge, however, then it is surely possible that you have forgotten the experience of naming the twinge. Do you need another "mark" that names the experience of naming the twinge? If every memory needs a name to help us remember it, then we seem to be caught in an infinite regress. Hobbes's use of marks seems to be like the old Indian theory that the world is supported on the back of an elephant. If the world

needs supporting, then the elephant needs supporting too. And if the elephant doesn't need support, then why does the world?

An **infinite regress** argument like this shows

a) that a proposed solution to a problem—in this case the problem of how the world stays in place—only creates another one—in this case, the problem of how the elephant stays in place, and
b) that every time we use the proposed solution to deal with the new problem there will automatically be yet another one to solve.

This shows that the proposed solution leads to the ridiculous position where we accept a strategy for solving a problem that creates a new problem for every problem it solves. In other words, it isn't a solution at all.

This infinite regress argument is the one that shows that there is no possibility in this case of checking that you are using the term "twinge" correctly. And, once this point is established, we have reached the heart of Wittgenstein's line of reasoning. Using the word "twinge" to refer to a private state involves conforming to the rule that you should say to yourself "twinge" only when you experience that private state. But the idea of trying to conform to a rule essentially involves the possibility that you might fail to apply it correctly, and in this case there is no such possibility. "Whatever is going to seem right to me is right. And that only means that here we can't talk about 'right.'" If we have mental states that are private, the argument shows that we can't talk about them, even to ourselves! Since it doesn't make sense to talk about such private states, Wittgenstein drew the conclusion that there could not be any: after all, if the sentence "There are private states" makes no sense, it certainly can't be true!

We might be able to turn the strategy of the infinite regress argument against Wittgenstein at this point, however. For the idea of a criterion of correctness is, presumably, the idea of some standard against which we can check whether we are following the rule properly. But isn't this the idea that we are applying the rule: check your use of the first rule against the standard? And if so, don't we

need a criterion of correctness to apply *this* second rule? Once this chain begins, there's no stopping it. So perhaps we shouldn't let it begin. Perhaps there can, in fact, be rules that we apply without criteria of correctness.

Actually, Wittgenstein himself pointed something like this out. For he argued that when we continue a numerical series (such as 1, 3, 5 . . .) it doesn't help to say that we are following a rule, because any way we go on conforms to some rule or other. So he seems to have concluded that it was just a fact that human beings presented with a series eventually just start to "go on in the same way."

Notice that these problems about following rules don't seem to have anything special to do with the point about privacy. If I had introduced the word "twingle" to refer to a kind of marble, there would need to be some criterion of correctness to decide whether I was using the word correctly. It would not be enough for me to say "Yes, a twingle" or "No, not a twingle" when each marble is shown to me: that could be like Mary's using the word "tonk." You would only be persuaded I was following a rule if there was something about each twingle—that it had more green than red in it, or that it was of a certain size, or something of the sort—that made me pick it from other marbles. It would not be satisfactory if "whatever was going to seem right to me was right."

Now, this may seem persuasive when it's applied to kinds of marble, but what about the concepts in terms of which you check my use of a rule like "Call it 'a twingle' only if it's green and large." What criterion of correctness is there for the use of the word "green" here? You could say the rule I'm following is:

G: Call it "green" only if it's green.

But if that will do as a criterion of correctness, why won't

T: Call it a "twinge" only if it's a twinge

do as a criterion of correctness in the original case? The difference between G and T seems only to be that G is a rule that other people can check that I am using correctly, whereas T isn't.

But that suggests that the problem of the mental twinge isn't so

much that I can't check on myself, but that other people can't check on me. And if that is what Wittgenstein thinks is the problem, then he seems to be begging the question. (An argument **begs the question** if it assumes what it sets out to prove.) For the private-language argument was meant to show that there couldn't be mental states that are knowable only by the person who has them; but now it looks as though that is one of the premises of the argument!

There has been a good deal of philosophical discussion about whether Wittgenstein was right to make his claim about rule following. As I have said, much of the first part of his *Philosophical Investigations* is concerned with an attempt to defend this claim. If it is right, this seems to be a very powerful argument against the Hobbesian view that the primary function of language is to help us remember our own experiences. So you might want to think about whether you should accept Wittgenstein's view that following a rule requires a criterion of correctness. If you do accept Wittgenstein's claim about rules, you have good reason to prefer behaviorism to Cartesianism. (Though it's worth insisting at this point that Wittgenstein himself did not endorse behaviorism.)

The behaviorist view of belief solves Descartes' problem: there is no difficulty for the behaviorist about the causal relations of mind and body. So the view has an answer to the mind-body question, namely, that having a mind is having a body with certain specific dispositions. And behaviorism certainly isn't open to the private-language argument. So it solves the other-minds problem because it says that we can know about other people's minds just as easily as we know about any dispositions. We can know about your pain just as easily as we can know that a glass is fragile.

But behaviorism seems to create new problems as it solves these old ones. Here is one of them. The behavior that most obviously displays belief is speech: if you want to know what I believe, the first step is to ask me. So, as I've said, some behaviorists have held that to believe something is to be disposed (in certain specific sorts of circumstances) to say certain sorts of words—the words, in fact that would ordinarily be taken to be the expression of that belief. The trouble is that this theory makes it impossible, for example, to explain the beliefs of nonspeaking creatures (including infants) and has led some philosophers to deny that such creatures can have

beliefs at all. Though there is something rather unsatisfactory about the privacy of the Cartesian mind, there is something simply crazy about the publicness of the behaviorist one. "Hello; you're fine. How am I?" says the behaviorist in a well-known cartoon, and the cartoonist has a point. We *do* know better than others about at least some aspects of our mental life. And the question for behaviorism is: why? It isn't just that we witness more of our actions than others. For in interpreting the minds of others we rely very much on their facial expressions; but we hardly ever see our own facial expressions at all. And, in fact, it seems obvious that I can tell what I am going to do next—what my current dispositions are—because I know (by, as it were, "looking inward") something of my own beliefs, desires and intentions.

Neither behaviorism nor Descartes' theory seems to be quite right.

1.4 Computers as models of the mind

In recent years, a new alternative to behaviorism has been suggested, which treats the mind neither as absurdly public, in the way behaviorism does, nor as completely private, in the way Cartesianism did. It is, in other words, a halfway house between behaviorism and Cartesianism, and it is called **functionalism**. Its recent appeal derives from the development of the very computers with which we began. For one way of expressing what functionalism claims is to say that it is the view that having a mind, for a body, is like having a program, for a machine.

A good way to start thinking about functionalist theories, however, is to look at similar theories of a simpler kind. Consider, then, what sort of theory you would need to give if you were trying to explain the workings not of something really complex, like a mind, but of something fairly simple and familiar, like a thermostat designed to keep the temperature above a certain level. What should a theory of such a thermostat say?

It should say, of course, that a thermostat is a device that turns a heater on and off in such a way as to keep the temperature above a certain level. Consider a thermostat that keeps the temperature above 60 degrees. An analysis of what something has to be like to do this job can be stated in a little theory of the thermostat.

> A thermostat has to have three working parts. The first, which is the heat sensor, has to have two states: in one state the heat sensor is ON, in the other it is OFF. It should be ON when the external temperature is below 60 degrees and OFF when it is above. It doesn't matter how the heat sensor is made. If it is a bimetallic strip, then maybe whether it is ON or OFF will depend on how bent the strip is; if it is a balloon of gas that expands and contracts as the temperature changes, then ON will be below a certain volume, OFF will be above. The second part is the switch, which needs to have two states also. It should go into the ON state if the heat sensor goes into its ON state and into its OFF state if the heat sensor goes OFF. Finally, we need the heat source, which should produce heat when the switch goes ON and stop producing heat when the switch goes OFF. (What I said about the heat sensor applies to the other parts too: it doesn't matter what they are made of as long as they do the job I have just described.)

This explanation of the nature of a thermostat also shows what a functionalist theory is, for this little theory is a functionalist theory. And what makes it functionalist is that it has all of the following characteristics:

> It says how a thermostat functions by saying:
> a) what external events in the world produce changes inside the system—here, changes in temperature cause the sensor to go ON and OFF;
> b) what internal events produce other internal events—here, changes from ON to OFF in the sensor produce changes from ON to OFF in the switch; and
> c) what internal events lead to changes in the external world—here changes from OFF to ON in the switch lead to increased heat-output; changes from ON to OFF produced reduced heat-output.

Anything at all that meets these specifications functions as a thermostat, and anything that has parts that play these roles can be said to have a heat sensor, a switch, and a heat source of the appropriate kind. In other words, at the most general level, a functionalist theory says what the internal states of a system are by fixing how they interact with **input**, and with other internal states, to produce **output**. What I mean by saying that the theory says what states *are,* can be explained by way of an example: our thermostat theory says what a heat sensor is by saying that it

a) changes from ON to OFF (and back again) as the external temperature falls below (and rises above) 60 degrees, and
b) causes changes that lead to an increase in heat-output if it is ON, and to a decrease when it is OFF.

A heat sensor is thus characterized by its **functional role**, which is the way it functions in mediating between input and output in interaction with other internal states. And we can say, in general, that a functionalist theory says what a state *is* by saying how it functions in the internal working of a system.

We can apply this general model to computers. They have large numbers of internal, usually electronic, states. Programming a computer involves linking up these states to each other and to the outside of the machine so that when you put some input into the machine, the internal states change in certain predictable ways, and sometimes these changes lead it to produce some output. So, in a simple case, you put in a string of symbols like "2 + 2 =" at a terminal, and the machine's internal states change in such a way that it outputs "4" at a printer. We can now see why computer programs can be thought of as functionalist theories of the computer. For a computer program is just a way of specifying how the internal states of the computer will be changed by inputting signals from disk or tape or from a keyboard, and how those changes in internal state will lead to output from the computer.

From one point of view—the engineer's—all that is going on in a computer is a series of electronic changes. From another—the programmer's—the machine is adding 2 and 2 to make 4. People who are functionalists about the mind—which is what I shall mean by "functionalists" from now on—believe that there are similarly two ways of looking at the mind-brain. The neurophysiologist's way, which is like the engineer's, sees the brain in terms of electrical currents or biochemical reactions. The psychologist's way, which is like the programmer's, sees the mind in terms of beliefs, thoughts, desires, and other mental states and events. Yet just as there is only one computer, with two levels of description, so, the functionalist claims, there is only one mind-brain, with its two levels of description. In fact, just as we can say what electrical events in a computer correspond to its adding numbers, a functionalist can claim that we

can find out which brain events correspond to which thoughts. Functionalism thus leads to monism. There is only one kind of thing, even though there are different levels of theory about it.

Functionalism starts with an analogy between computers and minds. It doesn't say that computers have minds. But if we go carefully through the functionalist's arguments, we will see how you might end up holding that they could have minds, even if they don't yet.

1.5 Why should there be a functionalist theory?

But before we look in more detail at some functionalist proposals, it will help if we consider why anyone should think that it ought to be possible to construct a functionalist theory.

In section 1.2 I raised two questions that a theory of the mind ought to answer: "What justifies our belief that other minds exist at all?" and "How are we to explain the relations of a mind and its body?" Functionalism answers the second question quite simply: a person's body is what has the states that function as his or her mind. Just as the physical parts that make up the "body" of the thermostat are what function as heat sensor, switch and heater, so the physical "hardware" of a computer is what has the states that function according to the program.

But consider now what functionalism implies in answer to the first question. To have a mind, functionalists claim, is to have internal states that function in a certain way, a way that determines how a person will react to input—in the form of sensations and perceptions. The answer to the other-minds problem must, therefore, be that we know about other minds because we have evidence that people have internal states that function in the right way. And, in fact, we do have such evidence, as the behaviorists pointed out. People with minds act in ways that are caused by what is going on in their minds, and what is going on in their minds is caused by things that happen around them. One reason for being a functionalist is, thus, that it allows you to deny the Cartesian claim that minds are essentially private, that only you can know what is going on in your mind. Wittgenstein's private-language argument gives us a reason for doubting that minds can be essentially private. We shall see in the next chapter why many philosophers have held that nothing that

exists can be knowable by only one person. For the thesis that there are things that cannot, even in principle, be known by anyone appears inconsistent with some very basic facts about knowledge. To make these arguments now, I would have to step ahead of this chapter's topic. But when you have read what I say in the next chapter (2.6) about **verificationism**, you might want to think again about whether functionalists are right in holding that it is an advantage of their theory that it denies that the mind is essentially private.

1.6 Functionalism: A first problem

So far what I have said about functionalism is very abstract. If we are to make it plausible, we will need a more concrete case to consider. Take beliefs.

Beliefs, for a functionalist, are characterized as states that are caused by sensations and perceptions of the appropriate kind, and that can cause other beliefs, and that interact with desires to produce action. Thus, for example, seeing a gray sky causes me to believe that the sky is gray, which may lead me to believe that it will rain, which may lead me to take my umbrella, because I desire not to get wet. Here the input is sensation and perception and the output is action; the internal states that mediate between the two are beliefs and desires.

There is an immediate and obvious problem for anyone who wants to say what beliefs *are* in a theory of this kind. Remember that a functionalist says what an internal state of the system is by describing its functional role: by saying how it functions in mediating between input and output in interaction with other internal states. Suppose we try to do this for some particular belief—say, the belief that the sky is gray. You might think you can say fairly precisely what would cause this belief. Looking up, eyes open, fully conscious, at a gray sky ought to do it. But the trouble is that this is really neither a necessary nor a sufficient condition for acquiring the belief. It isn't necessary, because you can acquire the belief in lots of other ways: looking at the sky's reflection in a pond, for example, or listening to a weather forecaster. It isn't sufficient, because, in suitably weird circumstances, you might reasonably believe that the sky wasn't gray when it looked gray. (Suppose, for example, I told you I had inserted gray contact lenses in your eye while you were asleep; suppose you

believed me. Then it would be very strange indeed if you came to believe the sky was gray when it looked gray.) The general point, so far as input goes, is that whether the evidence of your senses would lead you to some particular belief—here, that the sky is gray—depends on what else you believe.

A similar problem arises with output, though here the issue is even more complex. For what you do on the basis of the belief that the sky is gray depends not only on what other beliefs you have—for example, do you believe that gray skies "mean" rain?—but also on what desires you have—for example, do you want to avoid getting wet? So whereas for a heat sensor in a thermostat the effect of input doesn't depend on an indefinitely large number of other internal states, in the case of belief in a mind it does.

In finding a way to handle this increased complexity, the analogy with the computer is helpful. For, in this respect, computers are more like minds than like thermostats. The results of inputting a number to a computer depend also on a complex array of internal states. If I put in a "=" to an adding program after putting in "2" followed by "+" followed by "2", then the output will be "4"; but if I put in the same sign, "=", after putting in "4" followed by "+" followed by "2", then the output will be "6". Yet we can still give a functional role to each internal state of the system: we can do it by saying, for example, that when the adding program is in the functional state of having a "2" stored, entering "+" followed by any numeral, "n", followed by "=" will result in outputting the numeral "n + 2". The general strategy is this: we must specify the functional role of a state, A, by saying what will happen, for any input, if the computer is in state A, *but in a way that depends on what the other internal states are*.

So for a functionalist account of the belief that the sky is gray, we can say, at the level of input, that it will be caused by looking at gray skies, provided you don't believe that there's some reason why the sky should look gray when it isn't; and that it will also be caused by acquiring any other belief that you think is evidence that the sky is gray. And we can say, at the level of output, that having the belief will lead you to try to perform those actions that would best satisfy your desires—whatever they are—if the sky was in fact gray. Which actions you think those are will itself depend on your other beliefs.

It may look as though we have still not solved the problem we started out with. For this definition of the belief that the sky is gray still seems to define it in terms of other states of belief and desire, and these other states are ones we want to give functionalist definitions also. So, you might ask, isn't this sort of definition going to be circular? We are going to define the belief that the sky is gray partly in terms of what it will lead you to do if you believe that gray skies mean rain; but aren't we going to have to define the belief that gray skies mean rain partly in terms of what it will lead you to do when you believe the skies are gray?

This is a genuine problem if you want to use functionalist definitions, but there is a procedure that allows us to solve it in a way that avoids this circularity. Applying it in the case of beliefs is extremely complex, so it will help, once more, to start with a simpler case.

1.7 A simple-minded functionalist theory of pain

Pain is a mental state. Let's suppose we are trying to produce a functionalist theory of it. We begin by gathering together all the truths we normally suppose a mental state must satisfy if it is to be a pain. The American philosopher Ned Block has suggested how we might do it, for what he calls the "ridiculously simple theory," which we'll call "T", that

> T: "Pain is caused by pinpricks and causes worry and the emission of loud noises, and worry, in turn, causes brow wrinkling."

T *is* ridiculously simple. But we can still use it to elucidate some general points about functionalist theories of the mind. For with this simple theory we can see how the charge of circularity might be avoided.

So, begin with T. We write it as one sentence. Then, we replace every reference in the sentence to pain—whether actual or potential—by a letter, and each other, distinct, mental term by a different letter, to get

> T': X is caused by pinpricks and causes Y and the emission of loud noises, and Y, in turn, causes brow wrinkling.

(In this case, since there is only one other mental term, "worry," we only need the one extra letter, Y; but in other cases, as we'll see, we would need many more.) The next step is to write in front of this the words "There exists an X, and there exists a Y, and there exists a . . . which are such that" for as many letters as we introduced when we removed the mental terms. So, in this simple case, we get

R: There exists an X, and there exists a Y, which are such that X is caused by pinpricks and causes Y and the emission of loud noises, and Y, in turn, causes brow wrinkling.

Notice that we now have a sentence, R, that has no mental terms in it. It allows us to say how pain works without relying circularly on knowing what "worry" is. It would be circular to rely on our understanding of what "worry" is, because, in a full functionalist theory, we would be going on to define worry later. Now, finally, we can define what it is for someone to be in pain. For we can say that someone—let's call her Mary—is in pain if there exist states of Mary's, X, and Y, which are such that X is caused by pinpricks and causes Y and the emission of loud noises, and Y, in turn, causes brow wrinkling, and Mary has X. If Mary has such a state, a state that functions in this way, she is in pain.

Now, T is, as I said, ridiculously simple. But it has allowed us to see how to define one mental state—pain—that can only be explained in terms of its interactions with another mental state—worry—without assuming that we can define the other mental state first.

1.8 Ramsey's solution to the first problem

Now that we have seen how to solve the problem of defining one mental state without circularly assuming that we have already defined some others, let's see if we can see how to do this for belief. If we were to try to do this for belief, we should need many more letters than "X" and "Y." We call these letters "variables," and they function in a way I shall explain in the chapter on language. But the procedure would be exactly the same. We would first write down all the claims about beliefs and desires and evidence and action that we think have to be satisfied by a creature that has a mind. This body of

ideas is what is sometimes called our "**folk psychology**": it's the shared consensus of our culture about how minds work, the "theory" we learn as we grow up. If we join all the claims of folk psychology together with "and's" we will have one very long sentence, and that will be our functionalist theory of the mind. Call that sentence MT (for "mental theory"). From MT, we would then take out all the mental terms referring to beliefs and desires and replace them with "variables." The result of this we can call MT*. Finally, for each variable we should write "There exists a . . . " in front of MT*, and we would have a new sentence, which didn't have any mental terms in it. That sentence is called the Ramsey-sentence of the theory MT, because the British philosopher Frank Ramsey invented this procedure. The Ramsey-sentence of MT says, in effect, that something that has a mind has a large number of internal states—one for each variable—that interact with input and with each other in certain specific ways, to produce behavior. (I called the final version of the simple-minded theory of pain "R," because it's the Ramsey-sentence of the simple-minded theory of pain.)

In 1.4 I said that many philosophers who have thought about the other-minds question have wanted to be able to define mental states in such a way that it was always possible, at least in principle, that somebody else should know what is going on in your mind. Notice that this functionalist theory, set up in the way Ramsey suggested, seems to make this possible. For Ramsey's method allowed us to define pain in terms of its causes and effects, its functional role, in such a way that if we have evidence that someone's internal states would make them react in certain public ways—brow wrinkling and the emission of loud noises—in response to certain public events—pinpricks—we have evidence that they are in pain. It allowed us to do this without requiring that we know anything about the other internal states—in this case, worry—except that they too would have certain causes and effects, which could, in the end, be seen to show up in what people do. For the Ramsey-sentence of MT is true of someone if and only if he or she has a system of internal states that produces the right pattern of responses in output—in this case, brow wrinkling and loud noises—to input—in this case, pinpricks.

In the more complex case of beliefs, as we saw, we can proceed in a similar way. But here, just because the case is more complex and

there are so many more internal states, it may be very hard, in practice, to discover that the right complex pattern of dispositions to respond to input exists. So, while allowing us to take mental states seriously, functionalism also allows us to believe that they might be very difficult—indeed, practically impossible—for anyone, except perhaps the person who has them, to find out about. (I'll say something about how a functionalist might explain our knowledge of our own states later, in section 1.11.) It is in this sense that functionalism is a halfway house between Descartes and behaviorism. For Descartes, as we saw, left open the possibility that someone could have mental states that no one else could know existed even in principle. Functionalism denies this. Any evidence of the existence of the right (extremely complex) pattern of dispositions will be evidence of your mental states. For behaviorism, every mental state is nothing more than a disposition to respond to input. Functionalism denies this also. What someone with a certain belief will do when stimulated depends, the functionalist claims, on other internal states as well.

1.9 Functionalism: A second problem

I said, in 1.1, that from an epistemological point of view, it seemed plausible to say that M had a mind. We have been looking, in the last three sections, at functionalism about minds from an essentially epistemological point of view. We have seen that functionalism offers a plausible answer to the other-minds question: we can know, at least in principle, what is going on in other peoples' minds. But from the phenomenological point of view, which denied that machines could have minds, functionalism doesn't look so attractive. For if functionalism is right and to have a mind is to have certain internal states that function in a certain way, then anything that has states that function in the right way has a mind. That seems to have the consequence that if a computer had internal states that functioned in the right way, it would have a mind. And, the phenomenologist says, that is quite wrong. It isn't enough to have internal states that lead you to respond in the right way; you must also have an inner life. That inner life has to have the sort of character that Descartes thought it had. It has to be conscious mental life. And a machine could quite well behave in the right way without having any mental life at all.

If the phenomenologists are right, it follows that functionalism has failed to capture the essence of what it is to have a mind. For *if* they are right, a functionalist might say that a creature (or a machine) had a mind because it had internal states with the right functions, even though it did not, in fact, have a mind because it had no inner life. To understand this objection to functionalism, we must first try to make more precise what "having an inner life" means. The phenomenologist will usually explain this by saying that the difference between a creature with an inner life and one without an inner life is that there is *something that it feels like* to be a creature with an inner life, but nothing that it feels like to be a creature without one. If a person has an experience—say, seeing something red—we can ask what it feels like to have that experience. So, for example, if you, like me, are neither blind nor color-blind, then you know what it feels like to see red.

Suppose there was a machine that was sensitive to red things and had internal states that led it to say "That's red" and, generally, to do all the things that people do with visual information. The phenomenologist believes we could still not be sure that the machine knew what it felt like to see red. That is why the phenomenologist thinks that a functionalist might mistakenly think that a machine had a mind.

How are we to settle this dispute between the phenomenologist and the functionalist? It will help, I think, to consider it in the light of specific examples again; and, as we shall see, M and your mother provide just the right kinds of examples.

1.10 M again

M was a machine that would behave in every situation exactly like your mother. A machine that is made to have internal states that function like a human mind we can call **functionally equivalent** to a person. M and your mother are functionally equivalent. But phenomenologists might have different attitudes to them. The phenomenologist might say:

> How do I know whether M knows, as your mother does, what it feels like to see red? Your mother, I believe, does know, because she, like me, is a human being. I have reason to think that human beings with normal vision know what

> seeing red feels like. For I know what it is like, and I believe that other human beings are like me.

The functionalist replies:

> All the evidence you have that your mother knows what it is like to see red is from what she says and does. Since M does the same, it is unreasonable to believe that your mother has a mind and M does not.

Notice, first, that we cannot appeal to any evidence to settle the dispute. Even if we were discussing an actual machine instead of a hypothetical one, it wouldn't help, for example, to ask it if it knew what it felt like to see red. For any machine functionally equivalent to your mother would say "Yes" if you asked it if it knew what it felt like to see red, because that is what your mother would say. If you didn't believe that what the machine said was true, you might try to test it, just as you might try to test your mother, if you suspected that she was colorblind. But whatever she would do in the test the machine would do also. So no amount of such testing is going to give you a reason to say something about the machine that you wouldn't say about your mother. The phenomenologist's worry that M may lack mental states will never be settled by the kind of evidence that normally persuades us that people have them.

This is already a rather strange situation, since we normally think we can tell whether people know what it feels like, for example, to see red by testing their responses to red things. Nevertheless, despite the fact that no amount of evidence could settle the issue, the conviction that there is a real doubt about whether such a machine would have a mind is very widespread, including among philosophers. In the next chapter I shall be looking at arguments for the view that if no amount of evidence could decide an issue, there is no real issue. Someone who believes this is called a **verificationist**. And if verificationism is correct, then the phenomenologist must be wrong.

But even if the phenomenologist is right in thinking that some states, such as seeing that something is red, can be had only by someone with an inner life, there are other mental states for which this does not seem to be true.

Take beliefs once more. We do not normally talk of "knowing what it feels like to have a belief." Indeed, we can have beliefs—unconscious ones—that we are unaware of altogether, and even our conscious beliefs do not have a special "feel" to them. What does it feel like to believe consciously that the president is in Washington, or that the rain in Spain stays mainly on the plain?

If this is so, then, even if the phenomenologist was right to be suspicious about the claim that M knows what it feels like to see red, that would not give you a reason to doubt that it had beliefs. And, as the functionalist will insist, you would have all the same reasons for thinking that M did have beliefs as you have for thinking that your mother has them. But beliefs are a pretty important feature of people's minds, and if having beliefs is enough to have a mind, then, as I said, we might end up holding that machines could have minds, even if they don't yet.

1.11 Consciousness

The core of the dispute between functionalists and phenomenologists seems, then, to reside in their views of consciousness. Whether or not there are mental states—like unconscious beliefs—that are not in consciousness, there surely are conscious mental states. (If there are nonconscious mental states, then they will have to be picked out in some non-Cartesian manner. Since Descartes said that mental states were the contents of the conscious mind, for him the idea of an unconscious mental state would be a contradiction in terms.) What should the functionalist say is the characteristic feature of conscious mental states?

One possibility, which was proposed by the British philosopher Hugh Mellor (who happens to have been one of my own teachers), is to say that conscious states are the states of our own minds about which we currently have beliefs; they are the ones we are currently aware of. So, in particular, a conscious belief that it is raining will be present, on this account, when I believe that I currently believe that it is raining. Let's call a belief about your own current mental state a "second-order" belief. A conscious sensation (of redness, say) will occur when I have the belief that I am currently seeing red.

The functional role of these second-order states will be specified by saying that they are caused by first-order states—like seeing red

or believing it's raining—and that they play a role in shaping our behavior, in particular, in relation to ourselves. For one central form of behavior that a belief about something—call it "A"—can produce is behavior aimed at affecting A. So one kind of behavior my beliefs about my own current states is likely to affect is behavior aimed at changing or maintaining my current state.

An obvious example is this. I believe there's a reliable clock in the kitchen. I also want to know what the time is. So I go to the kitchen in the belief that if I look at the clock, I will come to believe that the time is whatever the clock says it is and that that will be (roughly) right. In order for this line of reasoning to work, however, at some point I have to be aware that I am uncertain of the time, and for that to happen, on the functionalist view, I have to have a second-order belief about my (current) mental state. It follows that, on the functionalist view, it is only if I am conscious of my ignorance of the time that you can explain why I go to the kitchen to look at the clock. So here is a kind of behavior that can only occur with consciousness.

On the other hand, if I am driving and a traffic light in front of me turns red, I can stop the car, as we say, "automatically": my belief that the red light is there and my desire to obey the traffic laws can operate directly without my coming to believe I believe anything. So, on this sort of functionalist view, some behavior can occur without consciousness.

There is another obvious kind of behavior that will require consciousness: telling you what I think or desire. For here, I need to form beliefs about my own mental states and then desire to communicate what I believe. Indeed, since, as we shall see in the chapter on language, communication is a matter of aiming to get people to believe things about your own beliefs, all communication will require second-order beliefs—beliefs about what I currently believe—and so will require consciousness.

The view that both going to find out what the time is and linguistic communication require consciousness is, I think, intuitively appealing, as is the view that we sometimes act on our beliefs without any conscious mediation. In fact, it seems reasonable to suppose that people can act not just without conscious mediation but when they are not conscious at all. Unconscious people—people when they are asleep, for example—can do things like swat mosquitoes.

An account of consciousness of this generally functionalist kind is likely to produce some impatience in the phenomenologist. For the apparatus of second-order states—states that are produced by other current states and that shape the behavior of a system by changing, or maintaining, its own mental states—could obviously be produced in an android: as I have already pointed out, M certainly has the full range of behavior that your mother has, including answering questions and going to see what the time is. Perhaps, the phenomenologist could concede, the functionalists' account of consciousness captures something about consciousness, just as their account of belief—with its role in shaping behavior—captures something about belief. But it leaves out entirely the phenomenological character of consciousness—what it feels like to be your mother or me or anyone else with consciousness. And without that character what you have is just a very good fake.

We seem to have reached an impasse: a situation where arguments have run out and there is still no secure conclusion. Faced with an impasse such as this, it is often helpful to ask whether there is some assumption shared by both parties to the debate—what we call a shared **presupposition**—that needs to be examined. If there are good arguments for both sides and both sides can't be right, maybe it's because they're both wrong in some way we haven't noticed. One shared assumption in the debate so far is an assumption about philosophical method. It is that we can discover the essence of the mind or of consciousness by a purely conceptual inquiry. We have been proceeding by making arguments that are based on our understanding of key terms, such as "belief," "behavior," "feeling." I have mentioned no experimental explorations of the nature of the mind by psychologists. (Indeed, I suggested at the start, you will recall, that it was irrelevant whether your mother had brain tissue as opposed to silicon chips in her head!) The only experiments I have considered are **thought experiments**, where you think about an imaginary case and ask yourself what you would say if it actually occurred. But you might object to this procedure on various grounds.

For one thing, it might matter whether the thought experiments were about things that could in fact actually happen. It is not at all obvious, for example, that there could in fact be a creature like M.

(Perhaps the only sort of thing that could exactly reproduce your mother's behavior would have to be made pretty much, molecule for molecule, as your mother is. And then most of us would probably suppose that there was something that it was like to be her, so that she would meet both the functionalist and the phenomenological criteria for being mentally the same as your mother.) What significance should we attach to our response to being told that something might happen, when, in fact, it can't happen? Why should we assume, that is, that ways of thought that work well enough in a rough-and-ready way in ordinary life would work just as well in a very different world?

Another, more fundamental line of objection would be to ask why we take it for granted that we have such internal states as beliefs and sensations at all. We are normally inclined to take it as obvious that someone has beliefs when they act, or sensations when they open their eyes on a lighted world. But the fact that this is part of the package of regular commonsense assumptions doesn't guarantee that we are right. People used to think it was obvious that some people were witches and that there were ghosts. (As a matter of fact, as we shall see in the final chapter, there are still places where most people think something similar.) Perhaps the very fact that our ordinary ways of thinking can lead both to functionalism and to phenomenology suggests that those ways of thinking are muddled. (After all, if you can draw incompatible conclusions from a set of assumptions, that shows there's *something* wrong with them!) Perhaps, in fact, we should rethink the sources of behavior.

The contemporary American philosopher Stephen Stich has suggested that we may indeed have to do just this. He has examined a good deal of recent work in **cognitive psychology**, the branch of the subject that seeks to explain how we perceive, remember, reason, decide, and then act, by postulating internal processes very like those in a computer program. Stich argues that there is already a good deal of evidence from cognitive psychology that our folk psychological theory is just plain wrong. In fact, he thinks, it may eventually turn out that there is simply nothing at all inside our heads that operates in the way that our folk psychology of belief and desire supposes. If that is true, then there would be no beliefs or desires! And then we should have to proceed, guided by cognitive psychol-

ogy or neuroscience (or perhaps some new field of science), to try to understand the causes of behavior in terms of internal states quite unlike those we have gotten used to. That is why the subtitle of his book *From Folk Psychology to Cognitive Science* is *The Case Against Belief*.

One natural response to this possibility is to say that even if science does end up showing this, we would still want to continue with our folk psychological theory for everyday purposes. We would still, that is, want to treat other people as if they had beliefs and desires and the rest, even if our official position was that they didn't. Another American philosopher, Daniel Dennett, has given this strategy a name: he calls it "adopting the **intentional stance**" toward them. We adopt the intentional stance toward someone (or something) when we predict its behavior on the basis of what it would do if it had beliefs, desires, and intentions, while leaving open the possibility that it does not, in fact, have them. Many of us already adopt the intentional stance toward objects that we don't believe have minds. It's perfectly natural to talk about what a computer "thinks," or to explain a chess-playing machine's moves by saying it's "trying to ward off my rook." But it's also perfectly natural to deny that any existing computer or chess machine really has beliefs, desires, or intentions. (Analogously, most of us still speak of the sun going "up" and "down" in the sky, even though we know that, strictly speaking, we're actually rotating around *it*.)

Stich argues that Dennett's proposal is intellectually irresponsible. What's the point of explaining the way people behave in terms of states they haven't got, once you develop a theory that explains how they behave in terms of states they have? But to this objection one might reply that there may be practical reasons why it is easier to use the folk psychological theory. Perhaps, for example, we are attached to this theory because it is programmed into us by evolution, so that, just as certain visual illusions persist, even once we know they are illusions, we will continue to think spontaneously of people as having beliefs, even once we realize they do not. Or perhaps the states that the new cognitive psychological theory postulates are rather difficult to identify, so that only a psychologist with special instrumentation can find out exactly what they are. (There is something odd about discussing what we should believe if there

aren't really any beliefs!) The rough-and-ready apparatus of folk psychology at least has the advantage that we can apply it pretty easily on the basis of looking and listening without special equipment.

But there is a natural response to both Stich's proposal and Dennett's, a response that challenges a presupposition they seem to share. It is that both of them ignore the fact from which the phenomenologist starts: the fact that each of us knows very well in our own case that we have beliefs, desires, sensations, and so on. In response to Stich, one wants to say:

> I grant that I might be wrong about how my mental states work, and about their causal relations. But I can't be wrong about whether I have mental states. They are, as Descartes rightly insisted, the one thing in the world I am most certain of. By "belief" I just mean something like the state I am in when I look at a vase and come to believe that it has a flower in it.

And to Dennett one might say:

> I can imagine taking the intentional stance toward somebody else, exactly because I can imagine that someone else doesn't really have beliefs and desires but only appears to do so. That is just the problem of other minds. But it's a problem of *other* minds; just because I have direct experience of my own internal states, I can't imagine taking the intentional stance toward myself.

1.12 The puzzle of the physical

I mentioned a little while ago that sometimes, in philosophy, it is important to examine the shared presuppositions of the parties to a debate, and I discussed a number of assumptions (some common to the functionalist and the phenomenologist, and one to Dennett and Stich) that might be questioned. I want to end this chapter by inviting you to think about another shared assumption: namely, that the puzzles about the relations between mind and body stem from the special character of the mind. After all, the idea that there is something special about the mind to be explained at all seems to presuppose that there is nothing much to be explained about the nonmental, the physical world. On the best current theories of nature, at one time the universe contained no minds, and they then evolved. One way of understanding how phenomenologists think about the mind-

body problem is to think of them as asking: "How could my mind—which I know from direct experience—be made out of matter, which seems so different from it?"

But why is it puzzling that minds are made out of matter? Stars, magnets, bacteria, and elephants are made out of matter, and each of these would have been hard to anticipate from the character of the universe before they emerged. We have learned about the properties of matter by seeing what can be made of it: we know that it is the kind of thing that magnets can be made out of, because we have found magnetic substances; we know that it is the kind of thing bacteria can be made out of, because we have found bacteria. Why is it especially hard to accept that it is the kind of thing minds can be made out of? Indeed, since the one thing of which each of us surely has the most extensive direct experience is our own mind, shouldn't we be puzzled, if we are puzzled by anything, by the nature of matter? How can it be, one might want to ask, that a world made of the sorts of things and governed by the sorts of laws that physicists now believe in should give rise to the astonishing range of experiences that each of us has every day?

1.13 Conclusion

In this chapter we have discussed some of the central questions of the philosophy of mind. We started by asking, "Can machines have minds?" But that led us to ask how we know that people have minds, and to think about the special kind of knowledge we seem to have of our own minds. Because we asked these epistemological questions, we came, at the end, to a point where we could go no further until we had thought more about knowledge. We were also led to consider what the relationship is between a mind and its body. And because causation seems very important to this relationship—because thoughts seem to cause actions, and events in the world seem to cause sensations—we found at another point that we could go no further until we had thought some more about causation. That is one reason why I haven't been able to settle the central dispute of this chapter—between the functionalist and the phenomenologist—decisively in favor of one or the other. But even if I had given an explanation of the nature of causation and of knowledge, I should not have been able to settle that question decisively. For it is a question

that divides philosophers now, and there is something to be said in favor of both sides. If, when we have gone further with knowledge, you decide to join the phenomenologist, on one hand, or the functionalist, on the other, I hope you will keep in mind that there are good arguments in support of each of them.

But I hope you will also entertain the possibility that these tensions in our thought reveal that we may need entirely new ways of thinking in order to understand what our brains are doing—even, perhaps, that we may end up giving up the idea of the mind altogether. After all, when Descartes began modern philosophy of mind, he did so by treating as a single category everything of which we can be directly conscious: but perceptions, beliefs, hopes, twinges, anxieties, emotions, wishes and desires—even as we normally think of them—are a fairly diverse bunch of things. Perhaps it was a mistake to think that a single theory that covered all of them could be constructed. And, I have suggested, perhaps it was also a mistake to think that the deep puzzle is about the nature of the mind, rather than about the nature of matter. If, after all, as the best current theories of nature suggest, minds appear in the world through evolution in material organisms, then one of the facts about matter that needs explaining is that it can produce all the many diverse phenomena that we call "the mind."