1

Dualism

You gotta have soul.

—Billy Joel

According to an ancient tradition, the mind is a nonphysical object. This doctrine is called **substance dualism**, and is the focus of the first half of this chapter (Sections 1.1 and 1.2). According to substance dualism, the mind is an entirely different sort of thing to the body. The body is a physical object—it's located in space; it's made from the atoms familiar to chemistry; it has a certain weight and height; and it can be seen and touched. The mind, on the other hand, is a nonphysical object. It's not located in space; it's not made from the atoms familiar to chemistry; it has neither weight nor height; and it can't be seen or touched. (In Chapter 8 we will refine our understanding of the difference between the physical and the nonphysical. For the moment we'll proceed on an intuitive understanding of the distinction.)

We've seen that, according to substance dualism, mind and body are different sorts of things or **substances**. (If it helps, read 'substance' as 'stuff'.) We can now see where the label 'substance dualism' comes from. According to substance dualism there are two distinct kinds of substances in the world: mental substances and physical substances. In other words, there is a *duality* of substances. Later in this chapter we will consider another form of dualism—**property dualism**. Whereas *substance* dualism claims that there are two fundamentally different kinds of *substances* in the world, *property* dualism claims that there are two fundamentally different kinds of *properties* in the world. (When philosophers use the word **property** they mean, roughly, 'feature'.) I'll say more about the distinction between properties and substances in Section 1.4.

Before getting started one brief terminological point is in order. Sometimes substance dualists call the nonphysical mind they postulate the 'soul'. However, when discussing substance dualism I'll tend to avoid the term 'soul' because of its associations with religious doctrines that are not part of substance dualism. For example, according to common usage the soul is an entity which survives the death of the body. However, the philosophical doctrine of substance dualism takes no stand on the afterlife one way or the other.

Imagine that, whilst on safari, Bloggs sees a lion a short distance away and runs back to his car. A few quick strides and he's safe inside. Here's how the substance dualist accounts for this series of events. First, light waves from the lion hit Bloggs's retina, stimulating it in a particular way. Bloggs's brain then extracts sensory information from the activation pattern on his retina, and passes that information on to his nonphysical mind. His mind interprets the sensory information it has received from the brain and recognizes that there is a lion present. It then decides that the best thing to do is to run quickly back to the vehicle. A message (RUN!) is sent from Bloggs's mind back to his brain. His brain sends the relevant signals to his leg muscles and he runs quickly back to the car.

According to substance dualism, mind and body, whilst quite distinct, interact with one another. Sensory information about the state of the world is sent from brain to mind, and decisions about how to react are sent from mind to brain. Your body is like a probe, sent by NASA to explore a distant planet. The probe sends pictures back to mission control, where scientists decide what the probe should do next. Instructions are then sent back to the probe which responds accordingly. The probe itself is entirely unintelligent. Similarly, information about the world is communicated by the body to the mind; the mind decides on a course of action and communicates the decision back to the body. The body itself makes no decisions.

It's important to note that the relations between the mind and the body are *causal relations*. The sensory information sent by the brain to the mind *causes* the mind to register the presence of the lion. And the mind's decision to run *causes* the brain to activate the relevant muscles. In other words, there are two-way causal interactions between the mind and the brain.

It's worth briefly considering two more examples.

1. Say that Bloggs burns his hand on the stove and, accordingly, feels a painful sensation. According to substance dualism, the damage to Bloggs's hand causes a message to be sent to his brain, which in turn sends a message to his nonphysical mind. The mind is then brought into a state which Bloggs recognizes as a painful sensation. According to substance dualism, experiences of pain are states of the nonphysical mind; the brain itself has no conscious experiences.
2. Say that Bloggs knows the following two things. (1) If it's Friday then it's payday. (2) It's Friday. From (1) and (2) he works out something else: (3) It's payday. According to substance dualism, all of these knowledge states are states of Bloggs's nonphysical mind. Moreover, his nonphysical mind's being in states (1) and (2) *caused* it to be in state (3). On this view, all rational inference occurs in the nonphysical mind; the brain is just plain dumb.

In this section we will consider four arguments in favor of substance dualism. The first three arguments all have the following structure:

1. Minds can _____.

2. No physical object can _____.

Therefore,

3. Minds are not physical objects.

Different arguments are obtained by filling in the empty slots in different ways.

1. *Could a physical object use language?* We obtain the first argument for substance dualism by filling in the empty slots with 'use language':

1a. Minds can use language.

2a. No physical object can use language.

Therefore,

3. Minds are not physical objects.

This argument was articulated by the seventeenth-century French philosopher, scientist, and mathematician, René Descartes (1596–1650). It seemed to him impossible that a physical object could generate and understand the rich variety of sentences which humans so effortlessly handle. Consequently, it seemed impossible to Descartes that the human mind could be a physical object.

Since Descartes' day, a great deal has been learned about language. In particular, we have come to appreciate that languages are regulated by a series of rules that specify which sequences of words count as grammatical sentences. These rules are called the **syntax** of the language. The syntax of English, for example, specifies that 'The boy ate the ice cream' is a grammatical sentence whereas 'Ate boy ice cream the the' is not. Syntax is *mechanical* in the sense that, in principle, a computer could be programmed to determine of any sequence of English words whether or not it's grammatical. I say 'in principle' because our understanding of syntax remains incomplete. Nevertheless, we have good reason to accept that a certain kind of physical object—a suitably programmed computer—could process the rules of language. Consequently, it seems that Descartes was wrong to at least this extent: a physical object could handle the syntax of language.

However, there is more to language than syntax. In particular, words and sentences have *meaning*. The ways in which meanings are assigned to the words and sentences of a language is called the **semantics** of that language. Recently, linguists and philosophers have began to unravel the mysteries of semantics. It's fair

to say that, at present, we don't have a fully worked out theory of semantics. But it's also fair to say that, at present, there seems to be little reason to doubt that a physical object could use language meaningfully. Descartes' argument from the claim that minds use language to the claim that the mind is a nonphysical object therefore seems mistaken.

2. *Could a physical object reason?* The second argument for substance dualism we will consider is very much like the first. Descartes not only doubted that a physical object could use language; he also doubted whether such an object could reason:

1b. Minds can engage in reasoning.

2b. No physical object can engage in reasoning.

Therefore,

3. Minds are not physical objects.

Descartes begins his defense of the crucial second premise by noting that reasoning is universal in this sense: there are many circumstances about which we can reason. He admits that there could be a mechanism for responding to any one circumstance (e.g. responding to dogs); however, he claims that there could not be a mechanism which responded to a multiplicity of circumstances (say, dogs, breakfast, and algebra). Consequently, a machine which could respond universally would require a vast number of mechanisms—one for each circumstance. But, he says, that's impossible: the number of mechanisms involved would be too great.

I'm unconvinced by Descartes' argument for the second premise. However, rather than directly discussing the second premise, I propose to briefly consider one kind of reasoning which modern machines can, at least to some extent, achieve—mathematical reasoning. (As a significant mathematician, Descartes would have been intrigued by the mechanization of mathematical reasoning.)

Just what do we mean by the expression 'mathematical reasoning'? If by 'mathematical reasoning' we mean something like 'the ability to correctly apply mathematical rules' then it's clear that physical objects *can* do mathematical reasoning. After all, the cheapest pocket calculator can apply the rules of addition, subtraction, multiplication, and so forth to a range of numbers.

'Mathematical reasoning' might, though, mean something else. It might refer to the ability to discover new mathematical truths and methods. Newton and Leibniz, for example, invented calculus—an entirely new way of solving certain mathematical problems. Could a computer be programmed to do mathematical reasoning in this sense? Could a computer discover calculus? This is a hard question, and one which we cannot address very fully here. What can be said is that *certain kinds* of mathematical discoveries can now be made by computers. These

discoveries involve deriving new mathematical truths ('theorems') from established mathematical claims ('axioms'). There are limits to how effective computers can be at making these sorts of discoveries. Nevertheless, it seems that at least some sorts of mathematical reasoning can be achieved by physical objects, and it is likely that future research will expand the range of mathematical problems which computers can solve.

3. *Could a physical object be conscious?* The third argument for substance dualism is as follows:

1c. Minds can be conscious.

2c. No physical object can be conscious.

Therefore,

3. Minds are not physical objects.

I suspect that considerations of consciousness weigh heavily with many dualists. Sometimes these considerations amount to little more than the bald intuition that no physical object could be conscious; sometimes they consist of sophisticated arguments. For the moment I propose to simply set aside the issue of consciousness. That issue is so important—and so difficult—that Part 4 of this book is devoted to discussing it. We will consider there whether the existence of consciousness provides good reason to endorse some form of dualism.

Before moving on to the final argument in favor of substance dualism, it is worth mentioning that each of the three arguments just discussed relies on Leibniz's *principle of the indiscernibility of identicals*. The German philosopher and mathematician Gottfried Leibniz (1646–1716) pointed out that if X and Y are identical then they have exactly the same properties. So, if there are properties of the mind which no physical object could have then, by the principle of the indiscernibility of identicals, the mind cannot be a physical object. And this is exactly the strategy adopted by the three arguments we have been considering.

4. *Doubt and existence*. The last argument for substance dualism which we will consider is also due to Descartes. In the *Meditations*, Descartes noticed that he could doubt the existence of his body. He begins by observing that sometimes when we dream we mistake our dreams for reality. For example, I might dream that I'm falling off a cliff, and whilst dreaming it seems to me entirely real that I'm falling off a cliff. Nevertheless, I'm actually asleep in bed. It follows that at least many of my present beliefs might be false. For example, it seems to me that at this moment I'm wide awake, sitting in front of my laptop. But it must be admitted that I could be asleep, dreaming that I'm sitting in front of my laptop. Consequently, my present belief that I'm sitting in front of my laptop can be called into doubt. Similarly, my

present belief that I have a body can be called into doubt. Perhaps I have no body but am presently dreaming that I do.

Descartes strengthened this line of thought by introducing a new thought experiment. It seems that I must admit that there might be an incredibly powerful alien determined to mislead me in all possible ways. This creature controls my thoughts, making me believe all sorts of things which are not true. Even if I admit that such a creature is possible, it seems that I must admit that my present belief that I have a body could be mistaken. Perhaps I am a disembodied spirit who has been deceived by the powerful alien into believing that I have a body.

Considerations like these led Descartes to the first premise of his argument:

(A) I can doubt that I have a body.

Next, Descartes took his thought experiments a little further. We have admitted that I might be dreaming that I'm sitting in front of my laptop. However, even if I'm dreaming, one thing remains certain: that I exist. My belief that I exist must be true, because even if I'm dreaming, I must exist in order to dream. Similarly, the alien might deceive me in all sorts of ways. Nevertheless, it remains certain that I exist. My belief that I exist must be true because, even if the alien is controlling my thoughts, I must exist in order to be controlled.

Considerations like these led Descartes to his second premise:

(B) I cannot doubt that I exist.

From (A) and (B) it seems to follow that:

(C) I am not my body.

We will return to the inference from (A) and (B) to (C) shortly. For the moment, notice that if we accept that I am my mind, then (C) entails the claim that:

(D) My mind is not my body.

Now (D) is not quite the same as substance dualism; nevertheless, establishing (D) would go a long way towards establishing substance dualism.

Let's now think about the inference from (A) and (B) to (C). At first glance, the inference from (A) and (B) to (C) would appear to have the same structure as this argument:

(A1) My car is red.

(B1) The car in front of me is not red.

Therefore,

(C1) The car in front of me is not mine.

The argument from (A1) and (B1) to (C1) is a good one. By the principle of the indiscernibility of identicals, if the car in front of me is my car it must have exactly the same properties as my car. Consequently, if my car is a different color to the one in front of me, then the car in front of me is not mine.

Now the argument from (A) and (B) to (C) also seems to rely on the principle of the indiscernibility of identicals. For it points out that I have one property—the property of it not being doubted that I exist—and my body has another property—the property of it being doubted that it exists. Since I and my body have different properties, it seems to follow that I am not my body.

But there's a catch. Consider the following argument.

(A2) I think my car is red.

(B2) I think the car in front of me is not red.

Therefore,

(C2) The car in front of me is not mine.

At first glance, this argument appears to rely on the principle of the indiscernibility of identicals. For it says that whilst my car has the property of being thought to be red, the car in front of me does not, and so the car in front of me is not mine. But it's clear that this argument does not work. Say that I have just won a blue car in a lottery, but mistakenly believe that I have won a red car. I go to pick up my new car and the lottery organizers show me a blue car. It really is my car, but I don't think that it is because I expect a red car. In that case premises (A2) and (B2) are both true: I think my car is red and I think the car in front of me is not red. Nevertheless, the conclusion (C2) is false: the car in front of me *is* mine.

More generally, the principle of the indiscernibility of identicals does not work when the properties in question involve psychological states like believing and thinking. Now this is crucial for the evaluation of Descartes' argument. For premises (A) and (B) both involve properties which involve the psychological state of *doubting*. Another example will make it quite clear that Descartes' argument doesn't work:

(A3) I can doubt I am the author of this book.

(B3) I cannot doubt that I exist.

Therefore,

(C3) I am not the author of this book.

Descartes has shown how I can doubt that I am the author of this book: I might have merely dreamed that I wrote it or my thoughts might be under the control of a powerful alien. And he has shown us how I cannot doubt that I exist. But it certainly does not follow that I am not the author of this book. Similarly, whilst I can doubt that I have a body and not doubt that I exist, it does not follow that I am not my body.

1.3 Arguments against substance dualism

In the previous section we considered four arguments in favor of substance dualism. None of these arguments was very convincing. In this section I will present three arguments *against* substance dualism.

1. *Princess Elizabeth's argument.* The substance dualist makes two claims about the mind. (1) Mind and body are radically different kinds of substances. (2) Mind and body causally interact. These two claims are in tension. If mind and body are supposed to be radically different, how can they causally interact? This objection was first put to Descartes by his contemporary, Princess Elizabeth of Bohemia (1618–80). Descartes' replies were highly evasive!

Princess Elizabeth's argument has a certain amount of force. Nevertheless, the argument can be overplayed. Notice that there are causal interactions between very different kinds of *physical* substances. For example, sunshine can heat metal, and yet sunshine and metal are quite different kinds of substance. The former is a kind of electromagnetic radiation; the latter an assembly of atoms. If quite different kinds of physical substances can interact, why can't physical and nonphysical substances interact? The crucial point, it seems to me, is not that mind and brain are (according to substance dualism) radically different kinds of stuff; rather, the crucial point is that the substance dualist has said absolutely nothing about the details of the interaction. Physics can tell us in considerable detail about the ways light affects matter, but the substance dualist can provide no details at all about the way the soul and brain affect each other.

2. *The explanatory completeness of physiology.* If you ask a physiologist to describe what happens when Bloggs runs away from a lion, they will say something like this. Running occurs when certain muscle groups—especially the muscles in the thigh—contract powerfully. The thigh muscles contract because they are stimulated by certain nerves. Those nerves arise in the spine, and are in turn stimulated by special spinal nerves. The spinal nerves in their turn are stimulated by the motor cortex—the part of the brain devoted to the initiation and control of movement. At this point the physiologist's account gets very complicated, but this much is clear. The motor cortex is stimulated by those parts of the brain responsible for decision making, which in turn receive input from the visual cortex—the part of the brain responsible for vision. (Remember that Bloggs ran away because he *saw* the lion.) And the activity in the visual cortex came about because Bloggs's retina was stimulated by the lion.

I have, of course, left out a great deal of detail. The sum total of what physiology has discovered about the causal background of even a simple movement would fill a dozen books. Nevertheless, it's clear that the theory offered by the physiologist is

a *physical* one. There has been no mention whatsoever of nonphysical substances. But if we can account for people's actions without appealing to nonphysical substances, then substance dualism is mistaken to at least this extent: the nonphysical mind does not cause people to behave as they do. Of course, the substance dualist could concede this point but still insist that the nonphysical mind is responsible for other aspects of our mental life. For example, it might be argued that, whilst not causally responsible for our actions, the nonphysical mind is nevertheless the seat of consciousness. We return to the issue of consciousness in Part 4. For the moment we can say this much: there is no need to believe in a nonphysical mind in order to explain action.

3. *The explanatory weakness of substance dualism.* In the Introduction we noted six general features of mental life which a good theory of mental states should be able to explain (or explain away):

1. Some mental states are caused by states of the world.
2. Some mental states cause actions.
3. Some mental states cause other mental states.
4. Some mental states are conscious.
5. Some mental states are about things in the world.
6. Some kinds of mental states are systematically correlated with certain kinds of brain states.

What is striking about substance dualism is the extent to which it fails to illuminate the items on this list. We have already seen that substance dualism has trouble explaining the first two items on the list. Moreover, it is completely silent about the third item: it says nothing at all about how one mental state causes another. How do states of nonphysical stuff bring about other states of nonphysical stuff? In particular, how is it that some of the causal relations between nonphysical states respect the canons of rationality? No answers are forthcoming.

Turning to the fourth item we can observe that substance dualists do not offer a *theory* of consciousness. They assert that nonphysical mental stuff is conscious; they do not tell us what it is about nonphysical stuff that facilitates consciousness. This problem is especially telling if we allow that some mental states are *unconscious*. What is the difference between conscious, nonphysical mental states and unconscious, nonphysical mental states?

Item (5) on the list of general features of mental states notes that at least some mental states are about things in the world: my belief that Mt Everest is 8,848 meters high is about Mt Everest. Theories of the 'aboutness' of mental states are called 'theories of **content**', and we discuss theories of content in Chapter 9. It is

not entirely out of the question that nonphysical states could be about things in the world; nevertheless, we don't at present have a dualist theory of content.

Finally, let's consider item (6). Why should states of a nonphysical mind be correlated with states of the physical brain? According to substance dualism, the brain plays a crucial role in mediating between the world and the nonphysical mind. Perceptual information about the world is conveyed to the mind via the brain, and instructions to move in certain ways are conveyed from the mind to the body via the brain. Consequently, the existence of correlations between mental states and brain states is not entirely unexpected. However, we know that damage to certain parts of the brain causes deficits of reasoning. In other words, we know that there are correlations between reasoning processes and certain brain states. According to substance dualism, though, reasoning occurs entirely in the soul. The correlations between reasoning processes and brain states are thus an embarrassment to substance dualism.

So far I have argued that substance dualism has little to say about the six items on our list. Moreover, there is little reason to expect that the explanatory situation will change. There simply are no obvious ways of developing nonphysicalist theories of perception, thought, action, or consciousness. In contrast, we shall see in later chapters that there are at least the beginnings of physicalist theories of most of the items on the list. Moreover, there are reasons to think that those physicalist theories might be developed in coming years.

The relative lack of explanatory power of substance dualism is, in my view, the most decisive reason available for discarding substance dualism. We should endorse the theory of mental states which most helps us understand the place of minds in the world, and substance dualism does very little to advance that understanding.

1.4 Property dualism

So far in this chapter we have largely been concerned with substance dualism. In this section I will briefly discuss an alternative kind of dualism—property dualism.

We haven't said very much yet about the distinction between substances and properties. For our purposes, a substance is something which could be the only thing in the universe. My body is therefore a substance, for we can easily imagine a universe which contains only my body. On the other hand, having a mass (roughly, weight) of 80 kg is not a substance, for we cannot imagine a universe which contains 80 kg *and nothing else*: there would have to be something else in the universe which had that mass. (This way of defining 'substance' is due to David Armstrong (1968: 7). I'm not entirely happy with it, but it will do for present purposes.)

We have seen that my body is a substance whereas having a mass of 80 kg is not. Having a mass of 80 kg is a *property*. Say that my body weighs 80 kg. Then one of my body's properties is having a mass of 80 kg. More generally: substances have properties.

Here are a few more examples. My car is a substance: we can imagine a universe which contains nothing but my car. One of my car's properties is being white. Another is having four tires. And a third is having the license plate 'UZR 155'. The Australian one-dollar coin in my pocket is a substance. It has various properties including being gold colored; being minted in 1998; and being in my pocket.

With the distinction between substance and property in place, we can now turn to the doctrine of property dualism. According to property dualism, mental states are nonphysical properties of the brain. The brain is a physical substance with various physical properties. For example, the typical human brain weighs about one kilogram; contains billions of neurons; has a blood supply; and so forth. That much is common ground. What's radical about property dualism is that it claims that, besides all of these physical properties, the brain has some *nonphysical* properties. These include being conscious; being in pain; believing that it is Monday; and wishing that it were Friday. In short, mental states are nonphysical properties of the brain.

There are various kinds of property dualism, but here we will focus on one especially important sort: *epiphenomenal* property dualism. Since 'epiphenomenal property dualism' is a bit of a mouthful, I will just say 'epiphenomenalism'. According to **epiphenomenalism**, physical properties of the brain cause nonphysical properties of the brain, but not vice versa. Consider again the example of seeing a lion (Section 1.1). According to epiphenomenalism, light waves from the lion stimulate Bloggs's retina in a certain way, and that in turn causes his brain to be activated in a certain way. In other words, his brain is caused to have a particular physical property—the property of being activated in a certain way. Bloggs's brain's having the physical property of being activated in that way causes it to have the nonphysical property of thinking 'LION!'

So far we have seen that, according to epiphenomenalism, mental states are nonphysical properties of the brain which are brought about by physical properties of the brain. The distinctive feature of epiphenomenalism is that the nonphysical properties of the brain do not, in turn, bring about physical states of the brain. Bloggs's 'LION!' thought has no causal powers—it doesn't *do* anything. But if his 'LION!' thought doesn't do anything, it does not cause him to run away. What, then, makes Bloggs run away when he sees a lion? According to epiphenomenalism, it is physical states of his brain alone which cause him to run away. So the full story according to epiphenomenalism is this. Light waves strike Bloggs's retina and cause his brain to be activated in a certain way. Call the physical property of having

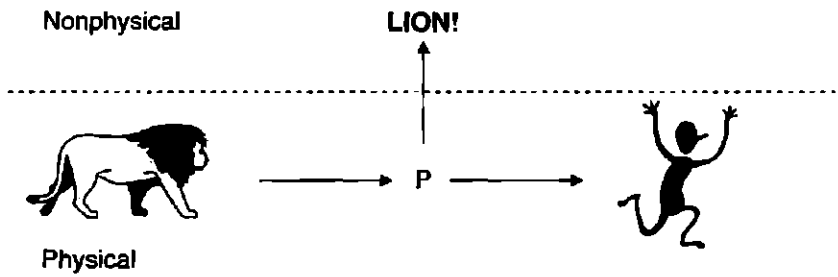

Figure 1.1 A diagrammatic representation of epiphenomenalism. The arrows represent the causal relation, with the arrowhead located at the effect

the brain activated in a certain way 'P'. P has two effects. First, it causes Bloggs's brain to have the nonphysical property of thinking 'LION!'. Second, it causes his legs to move so that he runs away. Figure 1.1 illustrates epiphenomenal property dualism.

It's important to stress that, according to epiphenomenalism, mental states are causally inert. My thought 'LION!' does nothing. What causes me to run away is a state of my brain.

1.5 Assessing epiphenomenalism

We saw in Section 1.3 that substance dualism faces three major difficulties: (i) Princess Elizabeth's problem; (ii) the **explanatory completeness of physiology**; and (iii) the explanatory weakness of substance dualism. Each of these problems also arises—in some form or other—for epiphenomenalism. Because the problems faced by epiphenomenalism overlap to a large extent the problems faced by substance dualism, my discussion of the former will be relatively brief. For more details, refer back to Section 1.3.

1. *Princess Elizabeth's problem.* Princess Elizabeth pointed out that there is a tension at the very heart of substance dualism: if mind and brain are radically different kinds of substance, how can they interact? A similar problem arises for epiphenomenalism: how can physical properties of the brain give rise to nonphysical properties of the brain? It must be admitted that this argument has a certain amount of force; however, since we allow causal interactions between quite different kinds of physical properties, why can't we allow causal interactions between physical and nonphysical properties? (For details, see Section 1.3.)

2. *The explanatory completeness of physiology.* When discussing substance dualism, we took note of the following difficulty. It's plausible that human actions like running away from a lion can be fully explained in terms of physical events like

muscle contractions and neuron discharges. But if every human action can be fully explained in terms of physical events, then it cannot be the case that nonphysical states play a crucial role in bringing about human actions.

Notice, though, that this difficulty does not arise for epiphenomenalism. According to epiphenomenalism, Bloggs's thought that there is a lion present is causally inert, and his running away from the lion is entirely due to activity in his brain. That is, epiphenomenalism is entirely compatible with the claim that physiology is explanatorily complete.

Epiphenomenalism, however, pays a high price for avoiding the objection from the explanatory completeness of physiology. For if mental properties are causally inert, we have to give up two of the general features of mental states which we noted in the Introduction:

(2) Some mental states cause actions.

(3) Some mental states cause other mental states.

(These were the second and third items in the list of general features of mental states given in the Introduction, hence the labels '(2)' and '(3)'.)

As the lion example makes clear, mental states do not, according to epiphenomenalism, cause actions. Consequently, accepting epiphenomenalism involves abandoning feature (2). Moreover, if mental states are causally inert, one mental state cannot cause another. Intuitively, we might think that Bloggs's LION! thought caused him to experience fear. However, according to epiphenomenalism, Bloggs's experience of fear was not caused by his LION! thought; rather it was caused by a physical property of his brain. Call the physical property of Bloggs's brain which caused the LION! thought 'P'. Then, according to epiphenomenalism, P also caused a further physical property of Bloggs's brain—call it 'R'—which in turn caused the nonphysical property of being afraid. (Figure 1.2 represents one way in which the details of this story might be filled in.) Consequently, accepting epiphenomenalism involves abandoning feature (3).

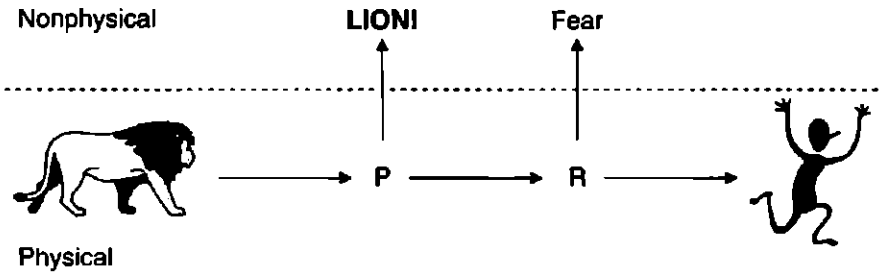

Figure 1.2 Epiphenomenalism. Note that the LION! thought doesn't cause the state of fear. Again, the arrows represent the causal relation, with the arrowhead located at the effect

Now it may be that our ordinary understanding of mental states is pretty much completely wrong and that we have to give up features (2) and (3). However, we would have to have very powerful arguments in favor of epiphenomenalism before it would be wise to give up so much of our ordinary understanding of mental states.

3. *The explanatory weakness of property dualism.* We saw in Section 1.3 that substance dualism explains very little about the mind. Moreover, it's not at all clear how substance dualism could be developed so that it began to illuminate the general features of the mind listed in the Introduction. Similar remarks apply to epiphenomenalism. Epiphenomenalism simply takes it for granted that physical properties of the brain can cause nonphysical properties of the brain, that mental states can be conscious, and that mental states can be about the world. Moreover, as we have just seen, epiphenomenalism *denies* that mental states cause action and that mental states cause other mental states.

I will bring this section to a close with a brief remark about consciousness and epiphenomenalism. We saw in Section 1.3 that substance dualism *takes it for granted* that some mental states are conscious; it does not *explain* how mental states could be conscious. There exists, however, a very powerful argument for the conclusion that consciousness is epiphenomenal. On this view, physical states of the brain give rise to nonphysical conscious properties which do not, in turn, cause anything. The argument, due to Frank Jackson, is discussed in Chapter 12.

1.6 Conclusion

In this chapter we have explored the idea that the mind is not physical. We have discovered that whilst the various arguments in favor of dualism are not especially convincing, the arguments against dualism are pretty powerful. In the next chapter we will consider one of the earliest physicalist theories of mental states—behaviorism.

SUMMARY

(1) Broadly speaking, there are two sorts of dualism—substance dualism and property dualism.

(2) According to substance dualism, mental states are states of a nonphysical object; according to property dualism, mental states are nonphysical properties of the (physical) brain.

(3) One way to defend substance dualism is to argue that there are things which the mind can do but which no physical object could do. We considered three examples of this style of argument. Two examples were unconvincing; assessment of the third, which concerned consciousness, was postponed until Chapter 12.

(4) Descartes offered an argument in support of substance dualism that was based on what can and cannot be doubted. However, his argument contains a serious error.

(5) One important version of property dualism is epiphenomenalism. According to epiphenomenalism, physical properties of the brain cause nonphysical mental properties, but not vice versa.

(6) Epiphenomenalism denies that mental states cause actions, and that one mental state can cause another mental state.

(7) The most significant difficulty for dualism in its various forms is its lack of explanatory power.

FURTHER READING

Churchland 1988: 7–22 provides a very elementary introduction to dualism. More advanced discussions are found in Armstrong 1968: Chs 2–4; Campbell 1984: Ch. 3; and Braddon-Mitchell and Jackson 1996: 3–13.

Descartes' concerns about language and reasoning are found in his *Discourse on the Method*, Part 5 (Descartes 1970: 41–2); for his argument based on doubt see his *Discourse on the Method*, Part 4 (Descartes 1970: 31–2). Princess Elizabeth's objection can be found in one of her letters to Descartes, dated 6–16 May 1643 (Descartes 1970: 274–5). (Note: Several good translations of Descartes' philosophical writings are available. Don't feel obliged to use the one to which I refer.) A good discussion of Descartes on substance dualism is Smith and Jones 1986: Ch. 3.

In Section 1.2 I mentioned contemporary theories of language. Pinker 1994 is a highly readable introduction to this fascinating area. In Section 1.3 I mentioned the possibility of providing a complete physical account of human movement. A nice introduction to the neuroscience of movement is Kosslyn and Koenig 1992: Ch. 7.

TUTORIAL QUESTIONS

(1) Describe substance dualism. (Use a picture if it helps.)

(2) What is Leibniz's principle of the indiscernibility of identicals?

(3) In your view, are there things which minds can do but physical objects could not achieve?

(4) What does it mean to say that physiology is explanatorily complete? How does the explanatory completeness of physiology pose a threat to substance dualism?

(5) How did Descartes establish that he can doubt the existence of his body?

(6) Describe property dualism.

(7) Describe epiphenomenalism.

(8) Give an argument against epiphenomenalism.

2

Behaviorism

Behave yourself.

—My mother

This chapter begins our exploration of physicalist theories of mental states by examining behaviorism. Two sorts of behaviorism will be discussed—**philosophical behaviorism** and **methodological behaviorism**. These two doctrines are closely related, although there is an important difference of focus. Philosophical behaviorism (also called 'logical' or 'analytic' behaviorism) offers a physicalist answer to the question, 'What are mental states?' In contrast, methodological behaviorism offers an account of how psychologists should go about their research. That is, methodological behaviorism proposes a *methodology* for doing psychological research. Despite these differences, both types of behaviorism emphasize the behavior people are disposed to produce under certain circumstances.

2.1 Philosophical behaviorism

According to philosophical behaviorism, mental states are **dispositions** (or 'tendencies') to behave in certain ways under certain circumstances. Pain, for example, is the tendency to cry or wince or . . . when you have broken your leg or burned your hand or . . . The first set of dots is intended to indicate that the behaviors associated with pain are not exhausted by crying and wincing—there are lots of things people do when they are in pain. Similarly, the second set of dots is intended to indicate that the circumstances associated with pain are not exhausted by broken legs and burnt hands—there are lots of painful **stimuli**.

According to philosophical behaviorism, to be in pain is to be disposed to do certain things when certain things happen to you. Here are a few more examples of philosophical behaviorist analyses of mental states. To believe that a lion is nearby is to run quickly to safety, or reach for your gun, or . . . when you see a lion, or hear a lion, or . . . Again the dots indicate that the lists of characteristic

behaviors and circumstances may be very long indeed. Another example: to be afraid of the dark is to scream or tremble or . . . when the light bulb fails or the candle blows out or . . .

It's important not to confuse philosophical behaviorism with two quite different claims. First, philosophical behaviorism does not claim that mental states are *the causes of* our dispositions to behave in certain ways under certain circumstances. According to philosophical behaviorism pain is the disposition to behave in certain ways when certain things happen to our bodies; it is not the cause of our disposition to behave in certain ways when certain things happen to our bodies.

Second, philosophical behaviorism must be distinguished from the claim that we *know about* the mental states of others by observing the way they react to the circumstances they are in. I might work out that Bloggs is afraid of the dark by noticing that he tends to scream or tremble or . . . when the light bulb fails or the candle blows out or . . . But claiming that that is how I work out what mental state Bloggs is in is quite different from claiming that his fear of the dark *is* his tendency to scream or tremble or . . . when the light bulb fails or the candle blows out or . . . (Compare: I might work out that there's a wildfire in the hills when I smell smoke, but that doesn't show that the wildfire *is* the smoke.)

When philosophical behaviorists use the term 'behavior', they are referring to physical events. Crying, wincing, running, reaching, screaming, trembling—these are all physical responses of the physical body. Similarly, behaviorists are only interested in the physical circumstances that trigger behavior. Breaking your leg, burning your hand, and seeing or hearing a lion are all physical events, as are the failure of a light bulb and the blowing out of a candle. It follows that philosophical behaviorism offers a physicalist account of mental states. According to philosophical behaviorism, mental states are dispositions to behave in certain ways under certain circumstances, and both the behavior and the circumstances that trigger it are understood to be physical events.

2.2 Arguments in favor of philosophical behaviorism

In the Introduction I gave a list of six features of mental states which a good theory of mental states should be able to explain. (I emphasized at the time that we may end up discarding one or more of the features on this list, but we would require good reasons for doing so.) One way to argue in favor of a theory of mental states is by showing that it is able to explain a number of these features. How well does philosophical behaviorism perform in this respect?

Philosophical behaviorism goes some way towards explaining three of the six features, and might—just might—have something to say about a fourth feature.

However, the two remaining features present a serious challenge to philosophical behaviorism. After briefly discussing the four features philosophical behaviorism can—or might—begin to explain, we will look in detail at two important arguments for philosophical behaviorism. (The two features philosophical behaviorism cannot explain will be discussed in the next section.)

The features of mental states which philosophical behaviorism goes some way towards illuminating are as follows. (I have retained the numbering used in the Introduction.)

1. *Some mental states are caused by states of the world.* Standing on a tack, for example, causes pain. Now, according to philosophical behaviorism, mental states are dispositions to behave in certain ways under certain circumstances. So, if philosophical behaviorism is to respect the first feature of mental states, it must be plausible that standing on a tack can make me disposed to say 'ouch', rub the sore spot, cry, and so forth. And surely that *is* plausible: when I stand on a tack I am disposed to do just those sorts of things.
2. *Some mental states cause actions.* Let's stick to the pain example. If philosophical behaviorism is to respect the second feature of mental states, it must be the case that my being disposed to say 'ouch', rub the sore spot, cry, and so on causes me to (for example) cry. And that's plausible. Consider a glass which is fragile. Something is fragile if it is disposed to break when dropped. If I drop the glass, one aspect of the cause of its breaking is its fragility. ('The antique glass broke when I dropped it *because* it was very fragile.') Similarly, part of the cause of my crying is that I was disposed to say 'ouch', rub the sore spot, cry, and so on. In other words if, as the philosophical behaviorist claims, pain is a disposition to cry (etc.), then one aspect of the cause of my crying is my being in pain.
5. *Some mental states are about things in the world.* Consider my belief that Mt Everest is 8,848 meters tall. That belief is about Mt Everest and represents Mt Everest as being 8,848 meters tall. In Chapter 9 we will look in detail at the issue of content. It is not entirely out of the question that a theory of content could be worked out within the framework of philosophical behaviorism. However, no one has yet provided the details of such a theory.
6. *Some kinds of mental states are systematically correlated with certain kinds of brain states.* Philosophical behaviorism respects the sixth feature of mental states. In the glass example, we said that the glass was disposed to break when it was dropped. Underpinning this disposition is a certain molecular structure. It's because the glass has that molecular structure that it broke when dropped. (The features of an object which underpin its dispositional properties are called the *categorical properties* of the object.) Now, plausibly, the features of the human body which underpin our behavioral dispositions are certain brain states. So

philosophical behaviorism is entirely consistent with the claim that mental states are systematically correlated with certain brain states.

I now turn to two important arguments in favor of philosophical behaviorism.

First argument. When someone wants a coffee they exhibit a certain behavioral disposition: they tend to drink coffee. And if someone often says that they want a coffee but never accepts one when it's offered, we're inclined to think that they don't really want a coffee. These observations illustrate an important point about mental states: there is a strong connection between mental states and dispositions to behave in ways characteristic of those mental states. Indeed, the connection is so strong that a person's persistent failure to exhibit the characteristic behavioral disposition of some mental state M is good evidence that they're not in mental state M.

How can the connection between mental states and behavioral dispositions be explained? If, as the philosophical behaviorist claims, to want a coffee is to be disposed to drink coffee, then it is no surprise that someone who wants a coffee tends to drink one. The connection between mental states and behavioral dispositions follows immediately from the philosophical behaviorist's analysis of mental states.

We can now sum up the first argument for philosophical behaviorism. There is a strong connection between mental states and behavior. Philosophical behaviorism can readily explain that connection since, according to philosophical behaviorism, mental states are behavioral dispositions. So the connection between mental states and behavior supports the claim that philosophical behaviorism is true.

There are, however, other theories of mental states which can explain the strong connection between mental states and behavioral dispositions. (We will look at one such theory in Chapter 4.) Consequently, the fact that philosophical behaviorism can explain the connection between mental states and behavioral dispositions isn't enough to establish that philosophical behaviorism is true. One of the *other* theories that can explain the connection may be true instead.

Second argument. In the 1920s and 1930s, a group of philosophers called the 'Vienna Circle' developed a new account of the meaning of a statement. A statement is a sentence which claims that the world is a certain way. 'The Eiffel tower is in Paris' and 'The moon is made of cheese' are both statements. The first makes a (true) claim about the location of a famous landmark; the second makes a (false) claim about the constitution of the moon. The theory of the meaning of statements advocated by the Vienna Circle is called **verificationism**. On this view, the meaning of any statement is its method of verification. Let me explain.

To verify a statement is to show that it is true (if it is true). Members of the Vienna Circle insisted that the only way to show that a statement is true is by

making *sensory* observations (that is, by looking, hearing, feeling, etc.). Let's take as our example the statement, 'The cat is on the mat'. That statement can be verified by *looking* for the cat; or *feeling* for the cat; or (I guess) *listening* for the cat. According to verificationism, then, 'The cat is on the mat' means 'If a normal observer looks in the right way they will have a cat-on-mat visual experience *and* if a normal observer feels in the right way they will have a cat-on-mat tactile experience *and* if a normal observer listens in the right way they will have a cat-on-mat auditory experience'.

To grasp the force of the verificationist theory of meaning, think about this. If I tell you that the cat is on the mat, what have I conveyed to you? Surely this: if you look in the right place you'll see that the cat is on the mat; and that if you touch in the right way you'll feel that the cat is on the mat; and if you listen in the right way you'll hear that the cat is on the mat; and so on. These considerations suggest that the meaning of a statement is its method of verification.

Statements which cannot be verified are, according to the Vienna Circle, meaningless. They thought that some statements made by earlier philosophers were meaningless because they could not be verified. For example, they rejected Descartes' statement that our minds are nonphysical objects because, since nonphysical objects cannot be seen, touched, smelled, heard or tasted, there is no way to verify Descartes' statement.

Now let's return to philosophical behaviorism. According to verificationism, the meaning of a statement is its method of verification. How would we verify a statement like 'Bloggs is in pain'? Well, we would note that Bloggs is crying or wincing or . . . after certain sorts of things have happened to his body. So according to verficationism, the meaning of 'Bloggs is in pain' is 'If a normal observer listens in the right way after certain things have happened to Bloggs's body they will have a Bloggs-is-crying auditory experience *or* if a normal observer looks in the right way after certain things have happened to Bloggs's body they will have a Bloggs-is-wincing visual experience *or* . . .'. But if that's what 'Bloggs is in pain' means, then pain must be the behavioral disposition to cry or wince or . . . when certain things have happened to our bodies. (Compare: if 'triangle' *means* 'three-sided figure' then a triangle *is* a three-sided figure.) So the verificationist theory of the meaning of statements leads quite quickly to philosophical behaviorism.

Most contemporary philosophers of language, however, no longer think that the meaning of a statement is its method of verification. The great American philosopher W. V. O. Quine (1908–2000), for example, thought that individual statements could not be verified; rather, entire theories comprising many individual statements are verified or rejected. Consequently, for Quine it is *whole theories* that have meaning; individual statements get their meaning only in virtue of being embedded in a much broader framework.

Important though Quine's ideas are, this is not the place to investigate them. For our purposes it is enough to say that the second argument for philosophical behaviorism rests on the verificationist theory of meaning, and that theory is almost universally rejected by contemporary philosophers.

2.3 Arguments against philosophical behaviorism

I remarked at the beginning of the previous section that there are two general features of mental states which present a very serious challenge to philosophical behaviorism. Those features are consciousness and causal relationships between mental states (again I retain the numbering from the Introduction):

3. *Some mental states cause other mental states.* For example, say that Bloggs has the following two beliefs:

   A. He believes that today is Friday.

   B. He believes that Friday is payday.

   These beliefs are likely to cause him to hold a further belief:

   C. He believes that today is payday.

   Notice that, besides the causal relationship between the first two beliefs and the last one, there is also an *evidential* relation between the first two beliefs and the last one. That is, the first two beliefs make it *reasonable* to believe the third. This is an example of the way in which our thought processes are often *rational*. Can philosophical behaviorism account for the rationality of our thought processes?

   In Chapter 6 we will look at one account of the rationality of thought—an account which takes the idea that the mind is a computer entirely literally. It is controversial whether the computational theory of the rationality of thought is the right theory. Nevertheless, two things are clear: (1) the computational theory of the rationality of thought is the only well-developed theory of rational thought we currently possess; (2) the computational theory is quite incompatible with philosophical behaviorism. Consequently, the fact that thought is often rational provides a major challenge to philosophical behaviorism: at present no behaviorist theory of the rationality of thought is available, nor is it clear how one could be developed.

4. *Some mental states are conscious.* In Part 4 we will examine the issue of consciousness in some detail. For the moment let us just note that philosophical behaviorism has nothing to say about consciousness. Say that I step on a tack and am immediately aware of a sharp pain in my foot. Now, according to philosophical behaviorism,

my pain is a disposition to behave in certain ways—to scream, wince, and so on. But it is utterly mysterious how my disposition to scream, wince, and so on could *hurt*. Why does my being disposed to act in certain ways feel like something? Isn't it possible that I could be disposed to scream and wince without actually *feeling* pain? Couldn't someone build a robot which has sensors to detect when it has stood on a tack, and which automatically makes a screaming noise whenever that occurs, but which has no feeling of pain?

I turn now to a pair of closely related arguments against philosophical behaviorism.

First argument. Imagine that Bloggs has decided to be the ultimate tough guy. When he stubs his toe he doesn't wince or cry or rub the sore spot; he just carries on as though nothing has happened. Even if he broke his leg he wouldn't scream or cry—he'd just calmly hobble to the nearest hospital. Of course, Bloggs still *feels* pain—it still hurts when he stubs his toe or breaks his leg—but he is no longer disposed to cry, wince, and so on.

Now imagine an entire community of people who, like Bloggs, have decided to become super-tough. In that community people still stub their toes and break their legs, and those members of the community who are unfortunate enough to stub their toe or break their leg still experience pain. Nevertheless, no one in that community is ever disposed to produce pain behavior—no one is ever disposed to cry or wince or scream.

This example shows that you can be in pain without being disposed to produce the kind of behavior typically associated with pain. Moreover, since no one in the community just described is inclined to produce pain behavior, the example shows that it can be perfectly *normal* for those in pain not to be disposed to produce pain behavior. (Indeed, anyone who did produce pain behavior would be considered very weird.) In other words, the example shows that being disposed to produce pain behavior is not **necessary** for being in pain. This point was first made by the contemporary American philosopher Hilary Putnam who coined the term 'superstoics' for people like our tough friend Bloggs. (See Putnam 1965.)

The superstoic example shows that being disposed to produce pain behavior is not necessary for being in pain. A similar example shows that being disposed to produce pain behavior is not **sufficient** for pain. Imagine someone who never felt pain. When they stub their toe it doesn't hurt; even if they broke their leg they wouldn't be in pain. For convenience we will call this person 'Smith'. As it happens, Smith is rather embarrassed about her condition, so she learns how to pretend to be in pain. When she stubs her toe she remembers to say 'ouch' and rub the sore spot. When she breaks her leg she screams and winces. Eventually, after a lot of practice, she learns to produce pain behavior indistinguishable from that of a normal person. Nevertheless, she never feels pain.

Smith is an example of a 'perfect pretender'. That we can coherently imagine a perfect pretender shows that a person can be disposed to produce pain behavior without actually being in pain. That is, it shows that being disposed to produce pain behavior is insufficient for being in pain.

Taken together, the superstoic and perfect pretender examples show that being disposed to produce pain behavior is neither necessary nor sufficient for pain. You can be in pain but not be disposed to produce pain behavior, and you can be disposed to produce pain behavior without being in pain. It follows that pain is not a disposition to behave in certain ways under certain conditions.

Second argument. Philosophical behaviorism assumes that for every mental state there is a corresponding set of behaviors. If you are in pain then you will do one or more of the following: cry, wince, scream, rub the sore spot . . . If you believe that there is a lion nearby you will do one or more of the following: run back to the vehicle, reach for your gun, call for help . . .

The superstoic example shows that this isn't true. When a superstoic is in pain he does not cry or wince or scream or rub the sore spot. He doesn't do those things because he wants to appear not to be in pain. Similarly, imagine that there is a lion nearby, and that you think that the best way to avoid being attacked by the lion is to stand perfectly still. In that case you wouldn't run back to the vehicle or reach for your gun—you'd stand very still.

These examples illustrate the point that how we react to our circumstances depends on our beliefs and desires. The superstoic's reaction to pain depends not just on the pain but also on his *desire* to appear not to be in pain. Similarly, how you react to a nearby lion will depend on your *beliefs* about lions. If you believe that the best way to avoid a lion is by standing very still, then you stand very still.

It's worth thinking a bit more about the lion example. Say that I believe that the best way to avoid a lion is to stand very still, but that I'm fed up with life and want to die. In that case I won't stand perfectly still because I believe that I won't get eaten if I stand still, and I desire to get eaten because I'm suicidal.

This example illustrates the complex relationships between mental states and behavior. It is rare that your behavior is determined by a single mental state; rather, how you behave is typically determined by a complex of mental states. Consequently, philosophical behaviorism is doomed. There is no set of behaviors which are characteristic of pain; what you do when you are in pain depends on what you believe and desire. And the same applies to every other mental state: what you do when you are in love, or want an ice cream, or believe in Santa Claus, depends on what else you feel, want, and believe. This fact about the relationship between mental states and behavior is a very important one. We will return to it in Chapter 4.

One final observation. Remember that whenever we gave an example of a philosophical behaviorist analysis of a mental state, we relied on a series of dots to show that the associated list of behaviors was incomplete. For example, we said that pain is the tendency to cry or wince or . . . under certain circumstances. We can now see that the list of behaviors is inevitably incomplete. How someone reacts to pain depends, as we have noted, not just on the pain itself but also on their other mental states. There are a great many mental states capable of influencing the way a person responds to pain, and different mental states will typically influence the pain response in different ways. Consequently, there are a very large number of possible pain responses. If I believe that the best way to relieve my pain is to jump in the air, I will (other things being equal) jump in the air; if I believe that the best way to relieve my pain is to walk backwards, I will (other things being equal) walk backwards; and so on (and on and on).

Our discussion of philosophical behaviorism is now complete. In the next three sections of this chapter we will examine methodological behaviorism.

2.4 What is methodological behaviorism?

The methodological behaviorist proposes that psychology restricts itself to seeking laws which link stimuli to behavior. 'Stimuli' includes both the sensory inputs which the organism is currently receiving and any relevant sensory inputs the organism has received in the past. Let's briefly look at an example.

A rat is placed in a cage which also contains a lever and a light. A pellet of food is released into the cage if, and only if, the bar is pressed when the light is on. As the rat wanders around the cage, it accidentally presses the lever when the light is on and receives a pellet of food. Quite quickly the rat's behavior is modified so that it persistently presses the lever when (and only when) the light is on. Ordinarily we would say that the rat has learned to get food by pressing the lever when the light is on. We will see shortly, though, that the methodological behaviorist will be disinclined to use everyday psychological terms like 'learn'.

By the end of the experiment, there's a correlation between the light going on (the 'stimulus') and the rat's pressing the bar (the 'operant'): the rat presses the bar when (and only when) the light is on. The correlation comes about because the experimental set-up links getting a food pellet (the 'reinforcer') with pressing the bar when the light is on. This is an example of what is called the *law of effect:* if an organism receives a reinforcer shortly after producing the operant in response to the stimulus, its tendency to produce the operant in response to the stimulus will increase. The law of effect is an example—indeed, it's the core example—of the sort of law the

methodological behaviorist is after: it describes a relationship between stimuli and behavior.

Notice that the law of effect makes no mention of the internal states of the organism. It does not say that the rat *learns* that it can get food by pressing the bar when the light is on, nor does it say that the rat *wants* food and *believes* that it can get it by pressing the bar. The methodological behaviorist insists that there is nothing to be gained by talking about the inner or psychological states of organisms. The best way to get on with psychology is to forget about what's in the mind and look for correlations between the inputs (stimuli) and outputs (behavior) of the mind.

In summary, methodological behaviorism instructs the psychologist to ignore the internal states of the mind and concentrate on seeing how organisms react to various stimuli. The aim is to find laws which relate stimuli to behavior. The laws will be of the form: if the organism receives stimuli S1, S2, S3, . . . then it will tend to respond with behavior B.

2.5 Arguments for methodological behaviorism

Methodological behaviorism advises the psychologist to avoid talking about mental states and concentrate on locating laws which link stimuli to behavior. A variety of arguments have been advanced in favor of this view. Here we will consider two.

First argument. The American methodological behaviorist B. F. Skinner (1904–90) insisted that it is bad science to theorize about unobservable states and properties. His concern was that, since such states and properties cannot be observed, we have no way of checking if our claims about them are true. Since science is only concerned with truths which can be established by good evidence, it should ignore claims about unobservable states and properties. (See Skinner 1980: 37–40.)

Now mental states cannot be directly observed. I cannot see your pains, nor can I see your belief that it's Thursday. Skinner concludes, therefore, that it's bad science to theorize about mental states. Consequently, he insists that psychologists should give up all talk about mental states.

The trouble with this line of argument is that pretty much all the best science deals with unobservables. The physicist can't see electrons; the paleontologist can't see dinosaurs (at best they can see the fossilized *remains* of dinosaurs); the geologist can't see the Earth's core. Nevertheless, our best theories in physics, paleontology, and geology talk about (respectively) electrons, dinosaurs, and the Earth's core.

One of Skinner's own examples is quite telling. He objected to the way in which early chemists tried to explain combustion by saying that a substance called 'phlogiston' is given off by burning objects. His worry was that phlogiston was

not supposed to be observable. We now know that the phlogiston theory of combustion is wrong. The great French chemist Antoine Lavoisier showed that combustion involves the interaction of oxygen with a flammable material. Lavoissier's theory is now universally accepted. But notice that oxygen is no more observable than phlogiston! The phlogiston theory wasn't rejected because it trafficked in unobservables; it was rejected because it was inconsistent with the experimental findings of Lavoissier and others.

Scientists routinely develop theories which posit unobservable states and properties. The theories are assessed by comparing the events that the theory predicts will occur with the events that actually occur. If a theory gets lots of predictions right—and doesn't get any predictions glaringly wrong—then we have grounds for thinking that the unobservables it posits actually exist. We will see shortly that an argument of this sort can be given in favor of the existence of mental states.

One final point. It might be argued that Skinner is wrong when he claims that mental states cannot be observed since we can all look inside ourselves and 'see' our own mental states. Skinner is aware of this move and rightly rejects it. A truly scientific psychology must rely on evidence which can be carefully checked. My reports about my own mental life cannot be carefully checked because no one else has that kind of access to my mental life. For all you know I might be lying when I say that I believe that it is Thursday, or I might suffer from a speech disorder which leads me to say words I don't mean.

Second argument. Previously we noted that if a theory gets lots of predictions right, and doesn't get any predictions glaringly wrong, then we have grounds for thinking that it is true. From the 1920s to the 1950s, methodological behaviorists were very successful at predicting a range of behaviors in a number of experimental animals (rats and pigeons were Skinner's favorites). Consequently, up until the 1950s, there were grounds for accepting methodological behaviorism. However, from the end of the 1950s onwards, it became increasingly clear that methodological behaviorism was of little value in human psychology. (We return to this point in the next section.) By the 1960s, the so-called 'cognitive revolution' was under way, with psychologists no longer wary of theorizing about mental states. Much of the best work currently being done in psychology makes unabashed reference to mental states.

2.6 Arguments against methodological behaviorism

We have already noted a powerful objection to methodological behaviorism: many of our best theories of human behavior make reference to mental states. In this section I will briefly note two further objections to methodological behaviorism.

First objection. In the example of the rat discussed above, the light going on was the stimulus and the pressing of the bar was the response. In a case like this we have no difficulty identifying the stimulus and the response. But when we turn to real-life human behavior it is typically much harder to identify the stimulus and the response. Consider the following situation, based on an example by the linguist Noam Chomsky.

You go to the art gallery with a friend and look at a painting by the Dutch master, Rembrandt. Your friend might say any of the following: 'Dutch'; 'Wow!'; 'It's a Rembrandt'; 'This old stuff really bores me'; 'Let's steal it'; 'Can you believe the City paid 32 million dollars for *that*?' The range of responses your friend might make to the Rembrandt is both very large and very diverse; consequently, there will be no law linking the stimulus (i.e. the Rembrandt) with a single response (or even with an easily identified set of responses). (See Chomsky 1959.)

In reply to this problem, Skinner is likely to claim that the Rembrandt is not a single stimulus. Rather, the Rembrandt is a large collection of stimuli, each of which elicits a different response. For example, it may be the way the paint is applied that prompts the response, 'It's a Rembrandt', whereas the amazing use of perspective prompts the response, 'Wow!' However, as Chomsky points out, the behaviorist has no way of predicting what the subject will say, nor of identifying in advance which aspect of the painting triggers which utterance (Chomsky 1959). When applied to cases like this, methodological behaviorism is empty. It amounts to nothing more than an unsupported assertion that every response is in fact under the control of some stimulus.

Second objection. Methodological behaviorism assumes without argument that the way we respond to every situation is entirely determined by our experiences. That assumption underpins the claim that we can predict how an organism will respond if we know what stimulation it is currently receiving and has received in the past. However, there is evidence that some aspects of our verbal responses are partly determined by innate knowledge—that is, by knowledge with which we are born. Many contemporary linguists (including Chomsky) think that we are born with knowledge of a 'deep' grammar common to all human languages. This is an extraordinary claim, and this is not the place to pursue it (see below under Further Reading for useful references). Note, though, that if we are in fact born with knowledge of some aspects of our world, our responses to the world are *not* entirely determined by our history of stimulation. Consequently, methodological behaviorism could be very wide of the mark indeed.

SUMMARY

(1) Broadly speaking there are two sorts of behaviorism—philosophical behaviorism and methodological behaviorism.

(2) Philosophical behaviorism answers the question, 'What are mental states?' According to philosophical behaviorism, mental states are dispositions to behave in certain ways under certain circumstances.

(3) Methodological behaviorism is a methodological stricture. According to methodological behaviorism, psychologists should restrict themselves to seeking laws which link stimuli to behavior.

(4) Historically, the most important argument for philosophical behaviorism is that based on the verificationist theory of meaning. However, the verificationist theory of meaning has largely been abandoned by philosophers of language.

(5) Taken together, Putnam's superstoic example and the related perfect pretender example show that pain behavior is neither necessary nor sufficient for pain.

(6) Methodological behaviorism was largely motivated by the mistaken idea that science should not traffic in unobservables.

(7) The existence of innate knowledge would seriously undermine methodological behaviorism.

(8) Chomsky pointed out that, in many cases of human behavior, there is no principled way of identifying the stimulus.

FURTHER READING

One of the most important presentations of philosophical behaviorism is Carl Hempel's 'The Logical Analysis of Psychology' (Hempel 1949). Hempel was strongly influenced by Rudolf Carnap's work in this area (see for example Carnap 1959). Gilbert Ryle's *The Concept of Mind* (Ryle 1949) is another important source. Hilary Putnam presents a devastating attack on philosophical behaviorism in his 'Brains and Behavior' (Putnam 1965). His superstoic example appears in that paper.

For good discussions of philosophical behaviorism see Campbell 1984: Ch. 4; Braddon-Mitchell and Jackson 1996: 29–38; and Kim 1996: Ch. 2.

The most important proponent of methodological behaviorism is B. F. Skinner. The most relevant of his copious works are *Science and Human Behavior* (Skinner 1953) and *Verbal Behavior* (Skinner 1957). Block 1980: Ch. 3 consists of key selections from *Science and Human Behavior*.

Famously, Noam Chomsky launched a devastating attack on methodological behaviorism in a review of Skinner's *Verbal Behavior* (Chomsky 1959). Chomsky's paper is rightly regarded as one of the most important publications in twentieth-century literature on the mind. An extract is reprinted in Block 1980 (Ch. 4). For a clear description of Chomsky's attack on Skinner see Bolton and Hill 1996: 7–10.

For a highly accessible account of the claim that some linguistic knowledge is innate see Pinker 1994. For a critique of that idea see Cowie 1999. (Unfortunately Cowie's book is rather hard.)

TUTORIAL QUESTIONS

(1) Describe philosophical behaviorism.

(2) Describe methodological behaviorism.

(3) What is the verificationist theory of meaning, and how can it be used to support philosophical behaviorism?

(4) Describe (i) the superstoic example and (ii) the perfect pretender example. Explain how these examples challenge philosophical behaviorism.

(5) Should science avoid postulating unobservable entities?

(6) What's wrong with saying that there must be *something* about the picture (or the picture plus the viewer's prior experiences) which disposed her to say 'Wow!'?

(7) How would the existence of innate knowledge challenge methodological behaviorism?

3

The identity theory

If you gave him a brain cell it'd be lonely.

—Old Australian insult

Very roughly, the **identity theory** asserts that the mind is the brain. More precisely, it claims that mental states are physical states of the brain. The qualification 'physical' is important. After all, property dualism asserts that mental states are properties of the brain (see Section 1.4). However, according to property dualism, mental states are *nonphysical* properties of the brain. Consequently, if the identity theory is to be distinct from property dualism, it must assert that mental states are *physical* states of the brain. For ease of expression, in this chapter I will simply say 'brain states' rather than 'physical states of the brain'. It is important to remember, though, that it is physical brain states that are being discussed.

The identity theory gets its name because it *identifies*—claims an identity between—mental states and certain brain states. I say 'certain' brain states because whilst the identity theory claims that every mental state is a brain state, it is not committed to the converse. In fact, it's certainly not the case that every brain state is a mental state. For example, in addition to billions of neurons, the human brain contains a large number of *glial* cells which play a supportive and protective role. It's unlikely that any mental state is identical with a state of one or more glial cells.

3.1 More about the identity theory

The idea that mental states are brain states is not new. The English philosopher Thomas Hobbes (1588–1679) and his French contemporary Pierre Gassendi (1592–1655) both made the claim more than three hundred years ago. However, the idea wasn't carefully expressed and defended until the 1950s when a group of Australian philosophers including J. J. C. Smart explored the idea.

These days the idea of mind-brain identity is commonplace. Indeed, it has crept into ordinary language with expressions like 'He's brainy' and 'I can't get my head around it'. However, when the idea was proposed back in the 1950s, it was

ridiculed. One English philosopher went so far as to suggest that Smart must have spent too much time in the hot Australian sun! (I owe this story to David Armstrong.)

When Smart articulated the identity theory he used a couple of analogies to convey his claim that mental states are brain states. According to Smart, mental states are brain states in the same way that water is H_2O and lightning is an atmospheric electrical discharge. These analogies are important for two reasons.

First, Smart's analogies are cases in which it took considerable scientific investigation to make the identifications. That water is H_2O cannot be established by casual observation, nor by thinking about the meanings of the terms 'water' and 'H_2O'. Similarly, the claim that mental states are brain states is not supposed to be an obvious truth which can be established by simple observation or by reflecting on the meanings of expressions like 'belief' and 'cortex'. (The cortex is a part of the human brain.) Rather, the claim that mental states are brain states is plausible in part because of advances in our understanding of the human brain.

In order to grasp the second reason why Smart's analogies are significant we need to understand the important distinction between **tokens** and **types**. Let's begin with an example.

Dingoes are a kind of wild dog found in many parts of the Australian outback. Say that we are camping in the outback and see four dingoes prowling around our campfire. In that case we have four tokens of the type *dingo*. The tokens are the individual animals; the type is the kind or class to which the individuals belong.

Notice that the four dingo tokens belong to a great many other types besides the type *dingo*. For example, they are tokens of the types *mammal, animal, material object,* and *scary thing which prowls around the campfire.*

Here's another example of the type/token distinction. On my bookshelf are two copies of Newton-Smith's nice book about the philosophy of science. One I bought for myself; the other was given to me by a friend. So on my shelf I have two tokens of the type *Newton-Smith's nice book about the philosophy of science.*

Now that we have in place the distinction between tokens and types, we can make a further distinction between **token identity** and **type identity**. Some examples will be useful. Posh Spice used to be a member of the British pop band *The Spice Girls.* After she left the band she married English soccer star David Beckham and now calls herself 'Victoria Beckham'. If you are invited to a party by Posh Spice you have simultaneously been invited to a party by Victoria Beckham. Posh Spice and Victoria Beckham are one and the same person. They are token identical. Similarly, the current President of the United States is George W. Bush. If you are invited to the White House by the current President you have simultaneously been invited to the White House by George W. Bush. George W. Bush and the current President of the United States are one and the same person. They are token identical.

In contrast, the identities between water and H_2O, and between lightning and atmospheric electrical discharge, are *type identities*. Every token of the type *water* is a token of the type H_2O, and every token of the type *lightning* is a token of the type *atmospheric electrical discharge*. Science has discovered that the type *water* and the type H_2O are identical, as are the types *lightning* and *atmospheric electrical discharge*.

We are now in a position to clarify the kind of identity which identity theorists want to assert between brain states and mental states. According to the identity theory, there is a type identity between mental states and brain states. For example, every token of the type *pain* is a token of the type *c-fiber firing*. Consequently, there is a type identity between pain and c-fiber firing. (I will often use the example 'pain is c-fiber firing' to illustrate the identity theory. This is a common practice in the philosophy of mind, but is not intended to be taken very seriously. There *are* nerve fibers called 'c-fibers' and they have something to do with painful sensations. However, it is unlikely that pain is identical to that particular type of neurological state. Moreover, whilst I will sometimes describe c-fiber firings as 'brain states', c-fibers are in fact peripheral nerves.)

Summing up, the identity theory asserts that every type of mental state is identical to a type of brain state. (It is not committed, though, to the converse.) The brain states in question are physical states of the brain. Moreover, the identities are not supposed to be discoverable by either simple observation or examining the meanings of the terms involved. Rather, they are analogous to scientific identities like 'water is H_2O'.

3.2 Arguments in favor of the identity theory

How well does the identity theory explain the six features of mental states noted in the Introduction? It's fair to say that the identity theory offers convincing explanations of three of the six features, and that it may turn out to be compatible with sophisticated attempts to explain two of the remaining features. However, one feature of mental states—consciousness—presents a serious challenge to the identity theory. In this section we will discuss those features of mental states which the identity theory, or a theory compatible with it, can explain. In the next section we will touch on, amongst other things, the issue of consciousness. A fuller discussion of consciousness will have to wait until Part 4. In what follows I have retained the numbering used in the Introduction.

1. *Some mental states are caused by states of the world.* Example: Bloggs's belief that there is a cup of coffee in front of him (mental state) is caused by there being a cup of coffee in front of him (state of the world).

If, as the identity theory claims, mental states are brain states, then the first feature amounts to the claim that some brain states (the ones held by the identity theory to be identical with certain mental states) are caused by states of the world. Research in neuroscience gives us grounds for thinking that this is true. For example, the causal impact of seeing a cup of coffee can be traced deep into the brain. Light from the cup stimulates the light-sensitive cells at the back of the eye (the retina), and information about the pattern of stimulation on the retina is carried into the brain by the optic nerve. (Intriguingly, the pattern of activation on the retina is reproduced many times in the visual centers of the brain.)

2. *Some mental states cause actions.* Example: Bloggs's desire for another coffee (mental state) together with his belief that there is more coffee in the kitchen (mental state), caused him to go into the kitchen (action).

If the identity theory is to explain the second feature of mental states, it must be the case that certain brain states cause actions like going to the kitchen for a coffee. Research in neuroscience makes it overwhelmingly likely that this is the case. We have very good evidence that actions are caused by activity in a part of the brain called the *motor cortex*.

3. *Some mental states cause other mental states.* Example: Bloggs's belief that it's Friday (mental state), together with his belief that Friday is payday (mental state), caused him to believe that it's payday (mental state).

If, as the identity theory insists, mental states are brain states, then the claim that some mental states cause other mental states is supported by the fact that some brain states cause other brain states. However, as we noted in the Introduction, there is something special about the way mental states interact with each other. Notice that my belief that it was Friday, together with my belief that Friday is payday, give me *good reason* to believe that it's payday. To put this point another way: the causal relations between mental states often respect the *rational* relations between them. In Chapter 6 we will look in a little detail at one theory of the rationality of thought. That theory is a **physicalist** one, and to that extent is compatible with the identity theory. However, it is controversial whether that account of the rationality of thought can be squared with the claim that mental states are brain states.

5. *Some mental states are about things in the world.* That is, they *represent* the world as being a certain way. For example, Bloggs's belief that Mt Everest is 8,848 meters tall is *about* Mt Everest and *represents* Mt Everest as being 8,848 meters tall. In Chapter 9 we will look at a range of theories of mental representation which are broadly compatible with the identity theory.

6. *Some kinds of mental states are systematically correlated with certain kinds of brain states.* According to the identity theory, mental states literally are brain states. Consequently, the identity theory smoothly explains the systematic correlation of mental states with brain states.

In the next section I briefly describe a historical case which strikingly illustrates the existence of mind-brain correlations.

3.3 Evidence from deficit studies

Deficit studies provide particularly striking evidence of mind-brain correlations. In a deficit study neuroscientists attempt to determine the function of a part of the brain by examining subjects who, due to brain damage, have lost a particular mental function. A great many mind-brain correlations have been explored in this way. In what follows I will sketch just one example to give the flavor of this research.

The 1840s was a period of great expansion of the American railway system. In those days construction teams relied on gunpowder to help clear away rock. A hole was drilled into the rock and a fuse inserted. The hole was then packed with gunpowder and the fuse lit. Everybody ran as fast and as far as possible before the gunpowder exploded. Finally, the rubble was cleared away by hand and the whole process repeated.

Phineas Gage was a highly responsible leader of a railway construction team. It was his job to carefully pack down the gunpowder before lighting the fuse—a process called 'tamping'. Gage had his own iron 'tamping rod' made. Now in a museum, it was just over a meter long and weighed around 6 kg. One end—the end inserted into the hole—was flat; the other pointed.

One day there was a terrible accident. It seems that Gage's tamping rod struck a spark from the wall of the hole, setting off the gunpowder prematurely. The rod, pointed end first, passed through Gage's left cheek and the front part of his brain (crucially, the prefrontal cortices), before exiting through the top of his skull. It was subsequently found some distance away. Incredibly, Gage survived. His personality was, however, drastically altered. Prior to the accident he was described as 'efficient and capable' (Damasio 1994: 4); after the accident he was careless and irresponsible. 'Gage', his friends observed, 'was no longer Gage' (Damasio 1994: 8). He could no longer hold down his job as team leader and began to drink heavily. He died in San Francisco at the age of thirty-eight.

The tragic case of Phineas Gage provides striking evidence of a correlation between a mental process—impulse control—and a part of the brain—the prefrontal cortices. Whilst Smart and his fellow identity theorists didn't know

enough about the brain to predict the details of that particular correlation, cases like Gage's provide important support for their view.

3.4 Arguments against the identity theory

There are two important ways of arguing against the identity theory. The first way appeals to Leibniz's principle of the indiscernibility of identicals; the second involves the distinction between type identity and token identity. As we might expect, consciousness is a difficult problem for the identity theory; it will be discussed in the context of Leibniz's principle of the indiscernibility of identicals.

1. *Arguments based on Leibniz's principle.* As we saw in Section 1.2, Leibniz's principle of the indiscernibility of identicals says that if X and Y are identical, then they have all their properties in common. Example: say that Sally is the tallest person in the room. In that case, if Sally has an IQ of 175, so does the tallest person in the room; and if the tallest person in the room rides a Harley Davidson, so does Sally.

The example just given involves a case of token identity: Sally and the tallest person in the room are one and the same *individual*. However, Leibniz's principle also applies to types. For example, the type *water* is identical to the type H_2O. So if water boils at 100 degrees Celsius, so does H_2O; and if H_2O conducts electricity, so does water. Similarly if, as the identity theorist claims, pain is c-fiber firing, then any property of pain is a property of c-fiber firing, and vice versa. Consequently, if we can locate a property of pain which is not a property of c-fiber firing, or a property of c-fiber firing which is not a property of pain, then we will have proven the identity theory false.

Various suggestions have been made of properties which pain has but c-fiber firing does not, or vice versa. For example, my pain has the property of being located in my foot, whereas my c-fiber firing does not; my pain is sharp but c-fiber firings are neither sharp nor dull; and my c-fiber firing has a frequency (say 20 firings per second) whereas my pain has no frequency. Since my pain and my c-fiber firings have different properties, they cannot be identical. Consequently, the identity theory is false.

Let's take each of these examples in turn.

(i) *My pain is in my foot but my c-fiber firing is not.* In reply, the identity theorist can insist that, strictly speaking, my pain is not in my foot. The brain state which is identical to my pain is in my head. Rather than talk about a pain in my foot we should talk about having a pain of the in-the-foot kind. One state of my brain—call it 'B1'—is identical to my having a pain of the in-the-foot kind; another state of my brain—call it 'B2'—is identical to my having a pain of the in-the-hand kind; and so on.

The identity theorist's reply gains in plausibility when we reflect on the phenomenon of phantom pains. Some unfortunate folk who have lost a body part continue to feel pain which they say is in the missing part. For example, it is not uncommon for people who have had a foot amputated to experience what they call a pain in their foot. These pains can be excruciating, and are very difficult to treat. Now it's clear that there is no pain located in their foot for the simple reason that they have no foot. Rather, they have a brain state of the kind we earlier called 'B1'—a brain state identical to having a pain of the in-the-foot kind.

(ii) *My pain is sharp but nothing in my brain is sharp.* This argument takes too literally the expression 'sharp' in 'sharp pain'. Clearly, the expression is metaphorical. To have a sharp pain is to have a pain which feels a certain way—it is not to have a knife-like pain. The identity theorist can say that pains of the sharp kind are identical to a certain kind of brain state, whereas pains of the throbbing kind are identical to a different kind of brain state.

(iii) *My c-fiber firings have a frequency but my pains do not.* In reply to this objection the identity theorist will simply assert that we have discovered (somewhat surprisingly) that pains have a frequency. Remember that the identity theorist offers scientific identities like 'lightning is an atmospheric electrical discharge' as examples of the kind of identity she has in mind. Now if the identity between lightning and atmospheric electrical discharge is correct, lightning has a *voltage*. I guess that to the modern mind that may not sound too surprising, but two hundred years ago someone would have been puzzled by that claim. Similarly, given the current state of understanding of psychology and neuroscience, it will strike many people as a bit odd to say that pain has a frequency. Nevertheless, science has discovered that it does.

There is one more application of Leibniz's principle which we should briefly consider. There is something that it is like to be in pain—*it hurts*. On the other hand, it is very hard to conceive how electrical activity in a nerve cell could hurt. As Colin McGinn put it, how could technicolor consciousness arise from gray brain matter (McGinn 1991: 1)? So, it seems that pains have a property—hurting—which no brain state could ever have. Consequently, pains cannot be identical to c-fiber firing.

It must be admitted that consciousness raises very serious difficulties for the identity theory. However, further discussion of consciousness will be deferred until Part 4.

2. *Type identity and token identity revisited.* We saw in Section 3.1 that the identity theory identifies mental state types with brain state types. The emphasis on type identity has, however, been challenged. There is a general consensus amongst contemporary philosophers of mind that the type identities proposed by the identity theory have to be either restricted or replaced with token identities. To get

a grip on the concern about the type identities proposed by the identity theory, we will consider a few examples.

Let's agree that, for the sake of argument, in humans pain is c-fiber firing. Now we can easily imagine animals with nervous systems quite different from our own; more specifically, we can imagine animals which don't have c-fibers. Let's agree, again for the sake of argument, that squid have nervous systems quite different from our own and lack c-fibers. (This isn't at all implausible. The squid brain *is* very different from our own. From an evolutionary perspective, humans and squid are only very distantly related. You have to go back a very long way to find a creature which was an ancestor of both ourselves and the squid.)

So far we have assumed only that squid don't have c-fibers. It seems quite likely, though, that they experience pain (or at least we have no very good reasons to doubt that they can be in pain). Consequently, the identity theory is in trouble: if squid can lack c-fibers but feel pain, then it cannot be the case that pain is identical to c-fiber firing. Another example will help reinforce the point.

Imagine a group of aliens whose brains (if you can call them that) are made up of silicon chips. They certainly don't have anything even remotely like c-fibers. Nevertheless, we can imagine that they feel pain when, for example, they stub their toe on the way into the teletransporter, or get a sore throat from repeatedly shouting, 'Exterminate all Earthlings'.

The examples we have just considered support the idea that pain is identical to different physical states in different kinds of creatures. Pain is said to be **multiply realized**: in different creatures pain is 'realized' in different ways. One way to respond to the multiple realizability of pain is to restrict the type identities to species:

Pain-in-humans is type identical to c-fiber firings.

Pain-in-squid is type identical to d-fiber firings.

Pain-in-aliens is type identical to activity in silicon chip E.

This list of type identities could, in principle, be extended indefinitely. I will call the resulting theory of mental states the *restricted type identity theory* to indicate that the type identities proposed are restricted to a given species.

However, it's quite likely that there are relevant differences *within a single species*. For example, I believe that the Eiffel Tower is in Paris. Chances are you do too. So we both have a token of the type *belief that the Eiffel Tower is in Paris* 'stored' in our heads. However, it's likely that the exact way in which my token of that belief is stored in my head differs slightly from the way in which your token of that belief is stored in your head. Whilst there's good reason to think that the coarse anatomy of your brain is very similar to mine, and that the mechanisms whereby information is stored in the brain are similar in both cases, it's unlikely that information about the location of the Eiffel Tower is stored in precisely the same

'place' in both brains. Exactly how a piece of information is stored seems to depend on the other information your brain has already soaked up, and your brain has no doubt soaked up different information from mine.

By way of analogy, think about the way information is stored on the hard drive of a computer. The exact pattern of storage on a hard drive depends on the information already stored on it. New information often ends up scattered around on unused portions of the drive. Consequently, even if copies of the same document are stored on computers of the same model, it's unlikely that the document will be stored in exactly the same way in both machines.

So once again we have an example of multiple realization: the way your belief about the Eiffel Tower is realized will probably differ slightly from the way my belief is realized. However, unlike the pain case discussed earlier, these multiple realizations occur within a single species. (I am assuming here that you're a human being!) These considerations have led some philosophers of mind to abandon even the restricted type identity theory. On their view, the most we can say is that each mental state *token* is identical to some brain state token. In other words, these philosophers endorse only the token identity of mental states with brain states. I will sometimes refer to this view as the 'token identity theory'.

I will not try to settle the dispute between those who advocate the restricted type identity theory and those who only advocate the token identity of mental states and brain states. (If you want to explore that issue, see Further reading, below.) It's enough for our purposes to note that the identity theory, *as originally stated*, is mistaken: there are no simple type identities between mental states and brain states.

3.5 Reductive and nonreductive physicalism

The term 'reduction' is used in a great many ways, and for some people is a term of abuse. Even in the philosophy of mind the term is used in at least two ways.

1. *Intertheoretic reduction.* Sometimes it is possible to show that one theory (the 'reduced' theory) can be derived from another (the 'reducing' theory). In that case an **intertheoretic reduction** has been achieved. Notice that the emphasis here is on *theories*—'intertheoretic' means 'between theories'. The example of intertheoretic reduction standardly given is the derivation of classical thermodynamics from the kinetic theory of gases. The former theory describes the behavior of gases in terms of their temperature, pressure, and volume; the latter describes the behavior of gases in terms of the kinetic energy and impacts of gas molecules. The derivation is achieved with the help of 'bridge laws' which identify the terms of one theory with those of another. For example, the pressure of a gas is identified with the mean (or 'average') kinetic energy of its gas molecules.

2. *Ontological reduction*. Sometimes it's possible to show that what appear to be two distinct kinds of phenomena are in fact the same kind of phenomena; that is, sometimes we can establish type identities (see Section 3.2). In that case we can say that one phenomenon has been **ontologically reduced** to another. The classic example is water and H_2O. Water is type identical to H_2O, and the discovery that water is H_2O facilitated the (ontological) reduction of water to H_2O. (Why has water been reduced to H_2O rather than vice versa? The general idea is that chemistry has the resources to deal with a much wider range of phenomena than does a science that is restricted to studying water. Consequently, chemistry is held to be the more 'basic' or 'fundamental' science.)

We have seen that Smart's version of the identity theory proposes type identities between mental states and brain states. That is, it asserts a series of ontological reductions between the kinds found in psychology and those found in brain science: the former are to be (ontologically) reduced to the latter. Moreover, the identity theorist asserts that, if we can locate the appropriate bridge laws, psychology will be intertheoretically reduced to neuroscience. Smart's kind of physicalism is therefore often called *reductive* physicalism.

In contrast, the position which I have called the 'token identity theory' is a kind of *nonreductive* physicalism. It denies that there are type identities between mental states and brain states, and so is opposed to ontological reduction. Moreover, it denies that there is any meaningful sense in which intertheoretic reduction could be achieved. Since mental states are, according to the token identity theory, multiply realized, there can be no simple bridge laws linking mental states with brain states.

You might like to keep the expressions 'reductive physicalism' and 'nonreductive reductive physicalism' in mind as you read around the topic: you're very likely to come across them.

3.6 Conclusion

The identity theory has a great many advantages but also some striking disadvantages. Is it possible to avoid what is problematic about the identity theory without losing what is valuable? In the next chapter we will examine **functionalism** which neatly sidesteps the issues raised by multiple realization whilst retaining many of the attractive features of the identity theory.

SUMMARY

(1) According to the identity theory, mental states are brain states.

(2) According to the identity theory, the identities between mental states and brain states are analogous to scientific identities (e.g. water = H_2O).

(3) Types are kinds of things; tokens are individual members of types. Example: Lassie is a token of the type *dog*.

(4) According to the identity theory, the identities between mental states and brain states are type identities.

(5) The identity theory accounts for a number of the features of mental states discussed in the Introduction. In particular, it predicts the existence of mind-brain correlations.

(6) The multiple realization of mental states creates a major difficulty for the identity theory. The restricted identity theory and the token identity theory were developed in response to multiple realization.

(7) The identity theory is a kind of reductive physicalism; the restricted identity theory and the token identity theory are kinds of nonreductive physicalism.

FURTHER READING

The most important contemporary source for the identity theory is Smart's 'Sensations and Brain Processes' (1959); see also Place 1956. Good discussions of the identity theory can be found in Armstrong 1968: Ch. 6, Sections I–IV; Churchland 1988: 26–35; Braddon-Mitchell and Jackson 1996: Ch. 6; and Kim 1996: Ch. 3. The papers in Part 2 of Block 1980 are both relevant and of outstanding quality; they are, however, all rather hard. For an excellent discussion of the issue of token identity versus restricted type identity see Braddon-Mitchell and Jackson 1996: 96–101.

For a good discussion of intertheoretic reduction see Churchland 1986: Section 7.2. For more on reductive and nonreductive physicalism see Kim 1996: Ch. 9.

For a fascinating account of Phineas Gage's case see Damasio (1994: Chs 1 and 2).

TUTORIAL QUESTIONS

(1) Explain the type/token distinction.

(2) Give examples of (i) token identities and (ii) type identities.

(3) Does the identity theory assert type or token identities between mental states and brain states?

(4) Which of the features of mental states given in the Introduction can the identity theory easily account for? Which does it struggle to account for?

(5) The Phineas Gage case is an example of a deficit study. Can you find another example of a deficit study which reveals a mind-brain correlation?

(6) What does it mean to say that mental states are multiply realized?

(7) How does multiple realization challenge the identity theory?

(8) Describe (i) intertheoretic reduction and (ii) ontological reduction.

4

Functionalism

> . . . one of the major theoretical developments of twentieth-century analytic philosophy.
>
> —Ned Block

Philosophy is a hard subject, and even amongst professional philosophers there are major disagreements. The philosophy of mind is no exception, and as yet there is no consensus about the nature of mental states. (This is not to say that there has been no *progress* on the issue: we are now much clearer on which answers are the *wrong ones*, and we have a keener sense of what problems need to be solved.) Whilst there isn't complete agreement about the nature of mental states, it's fair to say that functionalism plays a central role in contemporary philosophy of mind. Even those philosophers who reject functionalism agree that they need to explain in detail what's wrong with it.

4.1 Introducing functionalism

In the previous chapter we noticed that mental states can be multiply realized. In humans the state which realizes pain is (say) c-fiber firing; in squid it's (say) d-fiber firing. Multiple realization raises a puzzle: what do old Eight-legs and I have in common when we are both in pain? It can't be c-fiber firing because Eight-legs has no c-fibers (or so I will assume). And it can't be d-fiber firing because I have no d-fibers (or so I will assume). In virtue of what, then, is it true that Eight-legs and I are both in pain?

Functionalism provides an answer to this puzzle. According to functionalism, c-fiber firing *does the same job* in me as d-fiber firing does in Flipper. On this view, to be in pain is to have an internal state which does a certain job. Which job is that? Very roughly, an internal state does the 'pain job' if it is caused by bodily damage and causes us to say 'ouch' and rub the sore spot. So, according to functionalism, to be in pain is to have an internal state which is activated by bodily

damage and which causes us to say 'ouch' and rub the sore spot. More generally, according to functionalism, to be in (or have) mental state M is to have an internal state which does the 'M-job'.

Confused? Not to worry. Let's work through some analogies and a couple of examples. After that, I'm pretty sure you'll get the idea.

First analogy. Practically all cars have carburetors. A carburetor is a device which combines petrol with air and delivers the resulting mixture to the engine. In my car the carburetor is mainly made out of brass. (I drive an old Ford.) In more modern cars the carburetor is made out of a more sophisticated alloy. In the future, car manufacturers may make carburetors out of high-tech plastic. It doesn't matter what a carburetor is made out of as long as it can combine petrol with air and deliver the resulting mixture to the engine. That is, something is a carburetor because it does a certain job—mixing petrol with air and delivering the resulting mixture to the engine—not because it is made out of some particular material.

In summary, carburetors are multiply realized. What my carburetor has in common with yours is that they both perform the same job: they both combine petrol and air and deliver the resulting mixture to the engine. It is irrelevant that my carburetor is brass and yours some high-tech plastic. All that matters is that they get the job done.

Second analogy. An antibiotic is a substance which does a certain job: it kills disease-causing bacteria without doing serious harm to the patient. Penicillin kills disease-forming bacteria without doing undue harm to the patient; consequently it's an antibiotic. Erythromycin also kills disease-causing bacteria without doing serious harm to the patient; consequently it too is an antibiotic. However, penicillin and erythromycin have quite different chemical structures.

In summary, antibiotics are multiply realized. What penicillin and erythromycin have in common is that they both do the same job: they kill disease-causing bacteria without doing serious harm to the patient. It is irrelevant to their being antibiotics that penicillin and erythromycin have different chemical structures. All that matters is that they get the job done.

According to functionalism, mental states are in important ways like carburetors and antibiotics. What makes a carburetor a carburetor is that it does the 'carburetor job'; what makes an antibiotic an antibiotic is that it does the 'antibiotic job'. Similarly, what makes a mental state the particular state it is, is that it does the job associated with that mental state. Here are a few examples.

First example. Let's return to the case of pain. The doctrine of multiple realization says that pain can be realized in a variety of different ways. Functionalism explains the multiple realization of pain as follows. According to functionalism an organism

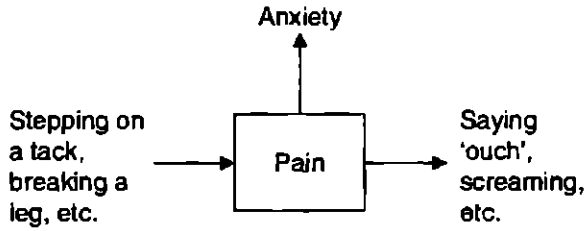

Figure 4.1 A highly simplified account of the pain role. The arrows represent the causal relation, with the arrowhead located at the effect

is in pain if it has a state inside it which does the pain job—or, as philosophers of mind prefer to say, if it has a state inside it which *occupies the pain role*. I'll say more about the pain role shortly—for the moment just think of it as the job pain does. Now in principle lots of different sorts of things could occupy the pain role, just as lots of different sorts of things can occupy the carburetor role or the antibiotic role. Consequently, pain is multiply realizable.

So what is the pain role? The pain role is defined in terms of *inputs*, *outputs*, and *internal connections*. The inputs are the circumstances which cause pain; they include stepping on a tack, breaking a leg, and burning your hand. The outputs are the behaviors which pain causes, including saying 'ouch', screaming, and rubbing the sore spot. The internal connections are the causal links between pain and other mental states. They include, for example, the causal link between pain and anxiety: pains (especially severe ones) often cause anxiety. (Figure 4.1 summarizes the pain role.) Putting all of this together, we can say that pain is a state which is caused by stepping on a tack (etc.), often makes us anxious, and causes us to say 'ouch' (etc.).

Second example. Consider my belief that a lion is near. (Let's assume that it's a wild lion.) On the input side my belief is caused by hearing a lion, or seeing a lion, or being told by a reliable witness that a lion is near. My belief has internal connections to, for example, fear: believing a lion is near very often causes fear. Things get more complex when we consider the output side. Typically, when we believe that there is a lion near we run away. That's because the belief that there is a lion near, together with the desire to live and the belief that the best way to escape is to run, causes running away. However, in combination with other beliefs and/or desires, my belief that there is a lion near may not cause me to run away. For example, imagine that Bloggs (foolishly) believes that the best way to escape from a lion is to stand perfectly still. In that case, his belief that there is a lion nearby, together with his desire to live, will cause him to stand perfectly still rather than run away. Again, imagine that Bloggs believes that there is a lion nearby and

believes that the best way to escape is to run away, but does not desire to live. In that case he may do nothing at all.

4.2 Functionalism and brain states

So far we've noted that, according to functionalism, mental states are the occupants of characteristic causal roles. In addition we've noted that, since in principle the roles characteristic of the various mental states could be occupied by a variety of different states, functionalism explains the multiple realizability of mental states. We turn now to the relationship between functionalism and (i) type identity theory; (ii) restricted type identity theory; and (iii) token identity theory. (For an explanation of the various kinds of identity theory see Section 3.4.)

Descartes had been dead for a couple of hundred years before functionalism was invented, so it's very hard to know what he would have thought of functionalism. But let's imagine that Descartes had not only thought of functionalism, but decided to accept it as an accurate account of the nature of mental states. Would he have had to give up substance dualism?

It is in fact possible to be a functionalist *and* a substance dualist. Consider pain. According to functionalism, an organism is in pain in virtue of having a state which occupies the pain role. Now it's conceivable that the pain role could be occupied by a state of a nonphysical substance. Consequently, it's conceivable that *functionalist* substance dualism is true.

Contemporary functionalists are, however, physicalists. They take it to be overwhelmingly likely that the characteristic causal roles of the various mental states are occupied by physical states of the brain. In other words, if functionalism is true then it is very likely that *some version* of the identity theory is true.

The contemporary Australian philosopher David Armstrong and the American philosopher David Lewis (1941–2001) independently struck on a very neat way of expressing these ideas. I will call the Armstrong/Lewis argument the *Transitivity Argument* because it relies on the logical principle of the **transitivity of identity**. Let's start with that principle.

Say that the tallest person in the room is identical to Sally, and that Sally is identical to the smartest person in the room. Then by the transitivity of identity we can conclude that the tallest person in the room is identical to the smartest person in the room. Using ' = ' for 'is identical to', we can express the principle of the transitivity of identity like this:

1. A = B.

2. B = C.

Therefore,

3. A = C.

Let's return to functionalism and take pain as our example. According to functionalism, pain is identical to the occupant of the pain role. Let's call the occupant of the pain role 'R'. Thus we arrive at our first premise:

1. Pain = R.

Now let's assume for the moment that R—the occupant of the pain role—is identical to c-fiber firing. Thus we have our second premise:

2. R = c-fiber firing.

By the principle of the transitivity of identity we can now obtain:

3. Pain = c-fiber firing.

In other words, if we assume that the occupant of the pain role is c-fiber firing, we can derive the type identity theory of mental states from functionalism.

However, as we saw in Chapter 3, pain is very likely to be multiply realized; for example it may be the case that whilst in humans pain is identical to c-fiber firing, in squid it is identical to d-fiber firing. Consequently, the assumption that R is identical to c-fiber firing is very probably mistaken. Far more plausible is the claim that *in humans* R is identical to c-fiber firing. Reconstructing our argument we get:

1. Pain = R.

2'. In humans, R = c-fiber firing.

Therefore,

3'. In humans, pain = c-fiber firing.

The conclusion expresses what in Chapter 3 we called the 'restricted identity theory'.

Finally, it may turn out that even the restricted identity theory is false. Perhaps the most that we can say is that in Bloggs R is identical to some brain state B. In that case we can derive the token identity theory from functionalism:

1. Pain = R.

2". In Bloggs, R = B.

Therefore,

3". In Bloggs, pain = B.

We have seen that from functionalism we can derive three versions of the identity theory: the type identity theory, the restricted type identity theory, and the token identity theory. The three derivations differ in that each relies on a different

second premise. In each case the second premise is an empirical claim—a claim that can only be established by observation and experiment. In the case of pain it's plausible that neuroscience will establish that the same type of brain state plays the pain role in all humans. However, it is likely that some mental states (for example, the belief that the Eiffel Tower is in Paris) are realized by subtly different brain states in different people.

4.3 Functionalism and the six features of mental states

In the previous section we saw that functionalism easily yields various versions of the identity theory. Consequently, functionalism's capacity to explain the six features of mental states identified in the Introduction closely parallels that of the identity theory.

1. *Some mental states are caused by states of the world.* We have seen that it is very likely that the states which occupy the functional roles characteristic of the various mental states are states of the brain. Consequently, for the functionalist the claim that some mental states are caused by states of the world is true only if some brain states are caused by states of the world. And, as we saw in Section 3.2, some brain states are indeed caused by states of the world.
2. *Some mental states cause actions.* Again recall that it is very likely that the states which occupy the functional roles characteristic of mental states are states of the brain. Consequently, for the functionalist the claim that some mental states cause actions is true only if some brain states cause actions. And, again as we saw in Section 3.2, some brain states do indeed cause actions.
3. *Some mental states cause other mental states.* If mental states are brain states, then the claim that some mental states cause other mental states amounts to the claim that some brain states cause other brain states. And that is certainly true. However, as we have noted in previous chapters, it's not merely the case that some mental states cause other mental states; in addition the causal relations between mental states sometimes mirror the rational relations between them. I'm getting a bit tired of the old examples, so here's a new one.

Say that Bloggs has a terrible hangover and that, whilst he can remember it's the weekend, he doesn't know which day of the weekend it is:

1. Bloggs believes that either it's Saturday or it's Sunday.

He then notices that he can't hear church bells, and realizes that it's not Sunday:

2. Bloggs believes that it's not Sunday.

Together, these two beliefs cause Bloggs to have a third belief:

3. Bloggs believes that it's Saturday.

Notice that in addition to the causal relation between Bloggs's beliefs, there is also a rational relation between them. (Strictly speaking there is a rational relation between the *contents* of Bloggs's beliefs.) The following is a valid argument:

1. Either it's Saturday or it's Sunday.

2'. It's not Sunday.

Therefore,

3'. It's Saturday.

So far we've just seen an example of the way in which the causal processes between mental states sometimes mirror the rational relations between them. Can functionalism explain that feature of mental states? The only detailed theory of this phenomenon we presently have—the computational theory—is in some important respects similar to functionalism. However, the computational theory insists that mental states are something more than the occupants of characteristic functional roles. In particular, it insists that they have a particular kind of *structure*. (We will develop this idea in Chapter 6.) If the computational theory is right, functionalism cannot be the whole story about mental states.

4. *Some mental states are conscious.* As usual, consciousness is a major headache. It seems that we can imagine a robot whose central computer has states which occupy the functional role characteristic of pain but which does not *feel* pain. If that's right, consciousness presents functionalism with a very serious problem.
5. *Some mental states are about things in the world.* In Chapter 9 we will see that there are, broadly speaking, two theoretical approaches to this issue. One of them—functional role semantics—sits very comfortably with functionalism. However the other approach, which includes the causal theory of content, requires at the very least additions to the basic functionalist framework.
6. *Some kinds of mental states are systematically correlated with certain kinds of brain states.* As we have noted, it's overwhelmingly plausible that the functional roles characteristic of the various mental states are occupied by brain states. Consequently, functionalism is compatible with the claim that there are systematic correlations between mental states and certain brain states.

Overall, the result is a mixed bag. Functionalism straightforwardly explains some of the six features; may succeed at explaining others; and struggles with the remainder.

We turn now to a pair of well-known antifunctionalist arguments.

4.4 Two famous arguments against functionalism

According to functionalism, mental states are the occupants of characteristic causal roles. This suggests two strategies for devising objections to functionalism. Consider some mental state M. If functionalism is true, any organism which is in M has a state which occupies the M-role, and any organism which has a state which occupies the M-role is in M. So, if we could find an organism that is in M but does not have a state which occupies the M-role, we would have shown functionalism to be false. Alternatively, if we could find an organism that has a state which occupies the M-role but which is not in M, we would have shown functionalism to be false.

The antifunctionalist arguments we will consider here all take the latter form: they all purport to describe a situation which, intuitively, involves no mental states but which is such that the relevant functional roles are occupied.

1. *The China Brain.* As we have seen, functionalists accept that, at least in principle, mental states could be realized by a wide range of physical—or even nonphysical—states. In the human case, mental states are most plausibly realized by brain states. We can, however, imagine them being realized by something quite different. For example, imagine that the entire population of China is enlisted to realize the mental states of a typical person—say Bloggs. The realization is achieved as follows. Each person in China is given a mobile phone and a set of instructions. The instructions tell them which numbers to ring when they have been rung by certain numbers.

For example, Jiang's instructions might be:

- If rung by 724 1144 then ring 722 9768 and 667 1849.
- If rung by 532 8181 and 95 5949 then ring 291 4245.

What Jiang is in fact doing is simulating the function of one of Bloggs's neurons—and this goes for every other person in China as well. Taken together, the population of China is simulating, neuron by neuron, Bloggs's brain. Consequently, whatever functional roles are occupied in Bloggs's brain are also occupied by the population of China. For example, if Bloggs believes that it is raining, the population of China believes that too. But that's absurd: a bunch of people ringing each other on mobile phones doesn't believe anything.

It's important to stress that when I say that according to functionalism the population of China believes that it is raining, I'm not referring to the beliefs of individual citizens. Rather, I'm referring to the entire population taken as a single unit. The point can be put this way. Say that there are a billion people in China, all of whom take part in the China Brain experiment. In that case, according to

functionalism there will be a billion *and one* minds in China. There will be a billion minds each of which belongs to exactly one Chinese citizen, and there will be, in addition, the mind realized by the entire population during the phone link-up.

There would, of course, be very many *practical* difficulties in actually setting up the China Brain experiment. For one thing, there are far more neurons in the human brain than there are people in China. In addition, we don't know anywhere near enough about the human brain to write out the instruction sheets for the participants. Nevertheless, functionalism is committed to the view that *if* such an experiment were undertaken, the population of China really would realize a mind.

The China Brain is supposed to be a case in which all the relevant functional roles are occupied but the corresponding mental states don't exist. For some people, the intuition that the China Brain has no mental states is very strong. But should we accept that intuition? Two factors would appear to drive the intuition. In my view, careful consideration of those factors reveals that the intuition based on those factors isn't worth much.

First factor: consciousness. My guess is that many people will doubt whether what it's like for the China Brain to believe that it's raining is the same as what it's like for Bloggs to so believe. Indeed, I suspect that most people will think that there is *nothing* that it is like for the China Brain to believe this or fear that. But we have already admitted that consciousness is a big problem for functionalism: the question is whether the China Brain presents a *further* problem to functionalism. We can put the issue this way: would the China Brain have a mind identical in all *nonconscious* aspects to Bloggs's mind? Will it process the same stimuli in the same way to yield the same output? Will its thoughts follow the same patterns? I suspect that for most people the answer will be 'yes'. In other words, what was driving their initial claim that the China Brain would not have a mind was a worry about consciousness, and we have already acknowledged that functionalism has a problem with consciousness.

Second factor: chauvinism. Chauvinism is a bias in favor of the familiar. Racism is a kind of chauvinism because it's a bias in favor of the race most familiar to the racist—his or her own. Now the mind realized by the China Brain would be a very different sort of mind to those with which we are presently most familiar. The minds with which we are presently most familiar are human minds, and human minds are found inside skulls and are realized by billions of brain cells which communicate with each other using special chemicals called 'neurotransmitters'. In contrast, the China Brain is not found in any single skull. It is distributed throughout a billion skulls which are widely located over a vast country. Moreover, the China Brain's 'neurons' (i.e. the individual Chinese citizens)

communicate with each other by mobile phones rather than by neurotransmitters. Consequently, there is a risk of chauvinism here—a risk of a bias in favor of minds realized in the way ours are realized.

Chauvinism about minds is nothing new. Europeans used to think that non-Europeans didn't have sophisticated minds. Such attitudes are now quite properly denounced as chauvinist. Similarly, some people have expressed chauvinism about animal minds, declaring that chimpanzees, for example, don't 'really' feel pain. But how do these cases differ from the China Brain? Isn't our rejection of the China Brain as mindless merely a chauvinistic refusal to accept that there might be minds realized in different ways to our own minds? Without an *argument* to show that the differences which exist between our minds and the China Brain's mind are significant, refusal to countenance the China Brain is just chauvinism.

In sum, the China Brain presents no *new* problems to functionalism. There is little reason to doubt that the China Brain's mind is identical in all nonconscious aspects to Bloggs's mind. Beyond that, it merely shows that, with a little bit of effort, we can create some pretty wild examples of multiple realization.

2. *Blockhead.* We are presented with choices every moment of our lives. Do I get up or stay in bed? Do I take a shower or a bath? Do I walk or take the bus? Usually we respond to a choice situation by *behaving* in some way: we stay in bed, or take a shower, or walk into town.

Now imagine that a scientist wants to build a robot which responds to every choice situation just as a typical human would respond. She begins by writing down all the circumstances the robot might find itself in: the alarm clock is ringing; in a café; in a burning building; on the Clapham bus; confronted by an enraged lion; and so on and on. (The list will be a very long one.) For each item on the list, the scientist thinks of a sensible response. So one small fragment of the list might look like this:

Circumstance	*Response*
The alarm clock is ringing.	Get up.
In a café at breakfast time.	Order breakfast.
In a burning building.	Find the fire escape.
On the Clapham bus.	Read a book.
Confronted by an enraged lion.	Run away.

The scientist now builds a robot which works as follows. First, the robot identifies the circumstances it is in. For example, it notes that it's in a burning building. It then searches through its list of possible circumstances until it finds the entry,

'In a burning building'. Next, the robot reads off the corresponding response, 'Find the fire escape'. Finally, it acts on that response—it looks for the fire escape. Since looking for the fire escape is exactly the sort of thing a typical human would do if they were in a burning building, the robot's behavior is just like that of a typical person.

The account of the robot I have just given is a bit rough. For one thing, descriptions like 'The alarm clock is ringing' and 'Order breakfast' are insufficiently precise. How a person responds to an alarm clock ringing depends on a number of factors including whether they are in bed or at an important meeting; the time of day the ringing takes place; and which day of the week it is. So the single entry, 'The alarm clock is ringing' needs to be replaced with a great many more specific entries with corresponding responses. (For example: The alarm clock rings on the morning of your exam → Get up.) Similarly, exactly how you order breakfast, and what you order, varies from place to place—there's probably not much point ordering kippers in central Mongolia. Consequently, the circumstance, 'In a café at breakfast time' needs to be refined, with each refinement matched to a refined response. (For example: In a café at breakfast time in central Mongolia → Order a glass of mare's milk.)

Second, the list of circumstances and responses needs to be carefully constructed so that the robot's responses are fairly consistent over time. People usually exhibit a degree of consistency in the responses they make to their circumstances: if a person has fried eggs, bacon, and sausages (with extra cholesterol) for breakfast, they're not likely to have a carrot sandwich (hold the butter) for lunch. Consequently, if the robot is to behave like a typical person, the list of circumstances and responses must exhibit an appropriate level of consistency.

The robot we have been discussing was first described by Ned Block (1981), and has since been called 'Blockhead' in his honor. In fact, Block's robot is a little different from the one just described as Block arranges the table of circumstances and responses into a branching structure called a 'look-up tree'. This technicality need not detain us; the robot as I have described it is enough to make Block's point. Let us turn now to the antifunctionalist argument Block makes with his Blockhead example.

Most people have the strong intuition that Blockhead has no mental states. That intuition is supported by the observation that Blockhead just blindly follows the instructions provided by the scientist. There is nothing going on inside Blockhead that looks remotely like deliberation. Blockhead no more has mental states than does a door bell which rings when you press a button. (Block himself remarked that Blockhead is no more intelligent than a toaster.) Block argues, however that functionalism is committed to the claim that Blockhead has mental states. If Block's right, functionalism is in big trouble.

Why might Block think that, according to functionalism, Blockhead has mental states? Consider what happens when Blockhead finds itself in a burning building. Seeing the flames causes Blockhead to search the list of circumstances for the entry, 'In a burning building'. Corresponding to that entry is the response, 'Find the fire escape', so Blockhead hurries around looking for the fire escape. Now, putting it crudely, functionalism says that Blockhead believes that it is in a burning building if it has an internal state caused by seeing flames and causing fire-escape-seeking behavior. And Blockhead does have such a state. As we have seen, the entry in the table of circumstances, 'In a burning building' is activated by seeing flames and causes fire-escape-seeking behavior. So, according to functionalism, Blockhead believes that it is in a burning building. But we have already seen that Blockhead has no mental states. So functionalism is false.

The trouble with Block's argument is that it misrepresents functionalism. When I sketched Block's argument I said that *putting it crudely* functionalism says that Blockhead believes that it is in a burning building if it has an internal state caused by seeing flames and causing fire-escape-seeking behavior. But that is to describe functionalism *far too crudely*. It's more accurate to say that, according to functionalism, the belief that the building is burning is a state which occupies a certain functional role. That role has inputs which include, but are not exhausted by, seeing flames; has outputs which include, but are not exhausted by, looking for the fire escape; and has internal connections to other mental states (for example to the belief that the situation is life threatening). Moreover, the outputs of the belief that the building is burning in part depend on the presence of other beliefs and desires. For example, the belief that the building is burning will only lead to fire-escape-seeking behavior in conjunction with the belief that the fire escape is the best way out of the building.

Once we articulate in a little bit of detail the functional role of the belief that the building is burning, it's clear that Blockhead has no such belief. Let's call the state in Blockhead which is caused by seeing flames and causes fire escape seeking the 'B-state'. The B-state would not cause fire extinguisher operating, nor would it cause 999 dialing, nor any of the other things people typically do when they believe the building is burning. Moreover, the B-state would not be caused by hearing the fire alarm or by being told by a reliable witness that the building's on fire. In addition, the B-state would not cause the belief that the situation is life-threatening, nor exhibit any of the other internal connections which the belief that the building is burning exhibits. And finally, the B-state's links to behavior do not involve any other mental states. In other words, the B-state does not occupy the functional role characteristic of believing that the building is burning, and so functionalism does not regard Blockhead as believing that the building is burning.

In sum, whilst the intuition that Blockhead has no mental states is very strong, that intuition is compatible with functionalism. Indeed, functionalism explains *why* Blockhead has no mental states.

4.5 Conclusion

Functionalism has made a very important contribution to our understanding of mental states. In particular it gives a beautiful account of multiple realization and allows us to understand much more clearly the relationship between mental states and brain states. Functionalism struggles to account for consciousness but—as we have seen—so does every other theory of mental states.

The real difficulty for functionalism lies, in my view, in explaining the rationality of thought. That's a theme to which we will return in Part 2.

SUMMARY

(1) According to functionalism, mental states are the occupants of characteristic causal roles.

(2) The causal roles of mental states are defined in terms of inputs, outputs, and connections to other mental states.

(3) Typically, a mental state causes behavior only in conjunction with other mental states.

(4) The Transitivity Argument has the following form:

(1) Mental state M = the occupant of causal role R.

(2) R = some brain state B.

Therefore,

(3) M = B.

Different versions of the identity theory are obtained by placing restrictions on the second premise.

(5) Functionalism readily accounts for some of the general features of mental states described in the Introduction. Whether functionalism can account for the remaining features remains an open question.

(6) Two standard objections to functionalism—the China Brain and the Blockhead—are not very convincing.

FURTHER READING

The classic early presentations of functionalism are Lewis 1966, Putnam 1967, and Armstrong 1968. Whilst these can all be recommended as marvelous examples of contemporary philosophical writing, Putnam's is probably the best place to start.

Excellent textbook presentations of functionalism can be found in Braddon-Mitchell and Jackson 1996, Chs 3 and 7, and Kim 1996, Ch. 5. Both books are quite a bit harder than this one.

What I have called the 'Transitivity Argument' was independently articulated by David Armstrong (1968) and David Lewis (1966, 1972, 1994). A more accessible discussion of the relationship between functionalism and the identity theory can be found in Braddon-Mitchell and Jackson 1996: Ch. 6.

Ned Block described both the China Brain and the Blockhead example in his important paper 'Troubles with Functionalism' (Block 1978). In that paper he also made significant distinctions between different types of functionalism, and discussed concerns about functionalism and consciousness.

TUTORIAL QUESTIONS

(1) Describe functionalism.

(2) In your view, which of the six features of mental states can functionalism handle?

(3) Sketch the Transitivity Argument, and show how functionalism is compatible with (i) the identity theory; (ii) the restricted identity theory; and (iii) the token identity theory.

(4) Describe the China Brain. Does it present a serious challenge to functionalism?

(5) Describe the Blockhead. Does it present a serious challenge to functionalism?

5

Eliminativism and fictionalism

I ain't got no use for what you loosely call the truth.

—Tina Turner

So far we've taken it for granted that mental states exist—that they're real. But what if mental states don't exist? What if they aren't real? Most of us are pretty confident that mental states are real, but it must be conceded that in the past people have been mistaken about the existence of all sorts of things. A thousand years ago there was widespread belief in the existence of dragons; now we know that dragons don't exist. A century ago it was believed that even 'empty' space was filled with a super-fine fluid called 'ether'. Now, thanks to Einstein, we know that ether doesn't exist. Couldn't a similar thing happen to our acceptance of mental states? Couldn't we come to reject mental states just as we have rejected dragons and ether?

Some philosophers think that we already have grounds for rejecting mental states. They think that mental states don't exist. Curiously, these philosophers are divided in their attitudes towards mental states. *Eliminativists* think that there are no mental states and it would be a good idea if we stopped kidding ourselves that there are. In contrast, *fictionalists* think that whilst there are no mental states, it's very useful to pretend that there are. We will return to this point towards the end of the chapter.

In order to understand **eliminativism** it's necessary to have a general grasp of the way in which theories give us access to reality. That's the topic of the next section.

5.1 From theory to reality

Why do we believe in atoms? After all, we can't *see* atoms in the way we can see bricks and books. In fact, even armed with the world's most powerful light microscope we can't see atoms. (Whilst images of atoms can be generated by electron microscopes, scientists were firmly convinced of the existence of atoms

long before electron microscopes were invented.) Our belief in atoms c
therefore, be based on direct sensory evidence. Rather, we believe in at
because our best theory of matter—atomic theory—says that there are atoms i
the world.

The atomic theory of matter says that material objects like tables, air, water, and planets are made up of atoms. Over one hundred different sorts of atoms (or *elements*) have been identified. Each element has different properties, and the properties of the elements determine the ways in which the atoms interact. (There are a few elements—the so-called 'noble gases'—which barely interact at all.) Scientists have been able to explain a great many of the properties of matter in terms of the interactions between atoms, and this information has allowed them to develop new, high-tech, materials.

Atomic theory has been extremely effective at predicting and explaining the properties of matter. Consequently, we have reason to think that it's *true*—or at least that it is a close approximation to the truth. If there really are atoms with the properties described by atomic theory, then matter will behave as atomic theory says it does. Since matter behaves as atomic theory says it does, we have good reason for thinking that there really are atoms with the properties described by atomic theory. Of course, we can't be *absolutely sure* that there are atoms; it could be a fluke that atomic theory accurately describes the behavior of matter. Nevertheless, the likelihood of such a fluke occurring is exceedingly low.

Time for a little jargon. A theory **quantifies over** something when it says that that thing exists. Atomic theory quantifies over atoms; Einstein's theory of special relativity quantifies over space-time but—as we saw earlier—it doesn't quantify over ether.

We can now sum up this section. The success of a theory gives us reason to believe in the existence of the things over which the theory quantifies. In particular, our best theory of some phenomenon provides us with good reason to believe in the existence of the things over which that theory quantifies. Good theories give us access to the way the world actually is.

5.2 Introducing eliminativism

In the previous section we noted how our best theories give us good reason to accept as real the things over which they quantify. The flip-side of this doctrine is that bad theories don't give us good reason to believe in the things they quantify over. Accordingly, if some theory T is the only grounds we have for believing in some entity E, and T turns out to be a bad theory, then we no longer have grounds for believing in E. When this happens we say that E has *been eliminated:* we used to

think that E existed, but it turned out that we were wrong and now we think that E doesn't exist. (Notice that eliminativism is *not* the doctrine that E used to exist but now it doesn't. Paleontologists are not eliminativists about dinosaurs; they merely think that dinosaurs are extinct.)

There are a couple of standard examples which are used to illustrate eliminativism. Let's quickly run through them before turning to eliminativism about mental states.

First example. 'Combustion' is the name given to the process of burning. An important eighteenth-century theory of combustion was the phlogiston theory. According to the phlogiston theory, burnable things (or 'fuel') contain phlogiston, and burning is the process whereby phlogiston is released from fuel. Things that aren't flammable—for example, bricks—contain no phlogiston.

The phlogiston theory has quite a bit of explanatory power. For example, with the addition of a further hypothesis it explains why sustained combustion requires a supply of fresh air. The additional hypothesis is that there is a limit to how much phlogiston a given volume of air can absorb. Once that limit is reached, no more phlogiston can be given off by the fuel, and so combustion ceases. A supply of fresh air sustains combustion by absorbing more and more phlogiston.

The phlogiston theory of combustion quantifies over phlogiston. For much of the eighteenth century, the phlogiston theory was the best theory available. Consequently, eighteenth-century scientists had good reason to believe in the existence of phlogiston. However, the brilliant French chemist Antoine Lavoisier proposed an alternative account of combustion: the oxygen theory. According to the oxygen theory, combustion is an interaction of oxygen and fuel. On this view, a supply of fresh air is needed to sustain combustion because the amount of oxygen in any given volume of air is limited. Once the available oxygen is used up, combustion ceases. A supply of fresh air sustains combustion by providing more and more oxygen.

Lavoisier's oxygen theory triumphed over the phlogiston theory because there was a striking fact about combustion which the oxygen theory could explain but the phlogiston theory could not. Somewhat surprisingly, the residue left over after combustion is complete *weighs more* than the original fuel. (Careful experiments are required to establish this result since the weight of any smoke released must be taken into account.) The increase in weight is very hard to explain on the phlogiston theory since, according to that theory, something is *given off* during combustion. On the other hand, the increase in weight is to be expected on the oxygen theory since, according to that theory, something is *absorbed* during combustion.

Scientists now universally accept that the phlogiston theory is false and that there is no such thing as phlogiston. In other words, phlogiston has been

eliminated. We used to think that there was such a thing as phlogiston; now we realize that there is not.

Second example. Human populations are subject to epidemics in which a disease sweeps through a community, often with fatal results. The great plagues which swept Europe in the Middle Ages are well-known examples of epidemics. People living at that time theorized about the origin of the plague. One very popular idea was that the plague was caused by witches—women who had thrown their lot in with the devil. Let's call this idea the 'witch theory of epidemics'. The witch theory quantifies over witches, and it supported the widespread belief that witches existed.

Very few people in the Western world would subscribe to the witch theory of epidemics today. Due to Joseph Lister, Ignaz Semmelweis, and others, the germ theory of epidemics is now universally endorsed in the West. According to the germ theory, epidemics are caused by the rapid transmission of microscopic organisms from one person to the next. In other words, germ theory quantifies over germs. (I'm using the expression 'germs' here to cover the whole range of microscopic pathogens, including viruses.)

The rise of the germ theory and consequent demise of the witch theory has led to the elimination of witches. We used to think that there were such things as witches; now we believe in germs instead.

5.3 Eliminativism about mental states

According to eliminativism, there are no such things as mental states. What motivates this extraordinary conclusion? To understand the eliminativist's argument, we must first understand the idea of **folk psychology**.

Practically everyone will tell you that agony is a kind of pain; that pains are unpleasant; that people who stand in front of a tree in good light will see the tree; that seeing generally leads to believing; and that love is very different from hate. They will tell you that people can remember some things about their past but not others; that if Sally wants to buy a book and believes that the bookshop is open, she will go to the bookshop; and that if Sally believes that it's Friday she will almost certainly believe that tomorrow is Saturday.

These are just a small sample of the very many obvious claims about the mind that are accepted by just about everyone. Such claims are sometimes called 'platitudes' about the mind. Taken together, the platitudes paint a highly detailed picture of the mental states and their interactions with each other and the environment; in other words, taken together the platitudes constitute an informal theory about the mind. That theory is called 'folk psychology' (sometimes

'commonsense psychology'). Folk psychology quantifies over a range of entities—beliefs, desires, pains, emotions, perceptions, and so on—and attributes various properties to those states. For example, it claims pains are unpleasant and that wanting to buy a book (together with other beliefs and desires) causes bookshop-going behavior.

This is where we return to eliminativism. According to eliminativism, folk psychology is 'radically false'; consequently, the states it posits—the mental states—don't exist. Just as the failure of the phlogiston theory gave us reason to turn eliminativist about phlogiston, and the failure of the witch theory of epidemics gave us reason to turn eliminativist about witches, the failure of folk psychology gives us reason to turn eliminativist about mental states.

Why, though, do the eliminativists think that folk psychology is a radically false theory? We will briefly explore three arguments offered by eliminativists against folk psychology. (These arguments are all from Churchland 1981: Section II.)

1. *Folk psychology is a 'stagnant research program'.* Scientific theories sometimes give rise to what are called 'scientific research programs'. A scientific research program consists of a number of scientists who share a common conception of what scientific problems need to be addressed, and how to address them. Newton's theories, for example, gave rise to a scientific research program which flourished for about two hundred years. It consisted of a number of scientists who applied Newton's theories to a large range of scientific problems. Research programs are said to be *progressive* when the scientists involved make a lot of progress; and they are said to be *stagnant* when the scientists fail to make significant progress. Stagnant programs are generally abandoned in favor of progressive ones, and are eventually forgotten by everyone except historians of science. (See Lakatos and Zahar 1978.)

Eliminativist Paul Churchland suggests that folk psychology is analogous to a scientific research program—a research program in which we are all engaged. And he suggests that it's a stagnant research program because it has made no progress—indeed, it has hardly changed—for centuries. Since folk psychology is a stagnant research program it's likely to be replaced by a more progressive one. In other words, folk psychology is likely to go the way of the witch theory of epidemics and the phlogiston theory of combustion. (In Churchland's opinion, neuroscience is likely to be the progressive research program which supplants folk psychology.)

Reply. It must be admitted that, in general, stagnation is evidence against a research program. So the crucial question is this: is folk psychology a stagnant research program? Churchland has urged that it is, but the issue is more complex than he makes out. To see why, we need to distinguish between folk psychology

and theories in scientific psychology which, whilst closely related to folk psychology, are nevertheless advances on folk psychology. Let me explain.

We have seen that folk psychology quantifies over a range of mental states including perceptions, sensations, emotions, and—importantly—propositional attitudes like beliefs and desires. Churchland's claim is that, since folk psychology is a stagnant research program, it's unlikely that these states exist. However, many theories in scientific psychology quantify over a similar range of states. Indeed, it's reasonable to suggest that scientific psychology has made important discoveries about the mental states originally posited by folk psychology. Here's an analogy. The ancient Greeks had ingenious arguments which showed that matter consisted of very tiny particles which they called 'atoms'. According to the Greeks, atoms were indivisible. However, we now know that atoms are *not* indivisible. In other words, modern physics has made important discoveries about the entities which the Greeks called 'atoms'. Similarly, modern psychology has made important discoveries about the entities originally posited by folk psychology. It has discovered, for example, that beliefs are not necessarily conscious. Consequently, whilst folk psychology itself may be a stagnant research program, it does not follow that the entities over which it quantifies don't exist since the very same entities are extensively discussed by highly progressive research programs in scientific psychology.

It will be helpful to have a label for those scientific psychological theories which quantify over states originally posited by folk psychology. For want of a better term I will, for the remainder of this chapter, use the term 'scientific folk psychology' for any such theory.

2. *Folk psychology fails to illuminate many important features of our mental lives.* Churchland draws attention to a wide range of topics about which folk psychology is largely silent. His list includes mental illness, creativity, sleep, vision, memory, and learning. These are important aspects of our cognitive lives, and any psychological theory which fails to contribute to our understanding of them is decidedly unattractive.

Reply. This argument is very similar to the previous one. In part folk psychology strikes us as stagnant because it fails to address the sorts of issues Churchland mentions. In replying to the previous argument we noted that whilst folk psychology itself may not have changed much for centuries, scientific folk psychology has made brisk progress. In particular, these sorts of theories have important things to say about many of the items on Churchland's list. I will briefly mention three examples.

First example: mental illness. According to an influential theory of depression, depressed people hold erroneous beliefs about themselves; in particular, they

believe that they are much less capable of dealing with life's difficulties than they really are. This has led to a form of therapy in which the therapist helps the patient identify and correct their erroneous self-beliefs. Interestingly, these forms of therapy are approximately as efficacious as drug therapies. For our purposes what is important is that this theory of depression quantifies over states which are quite recognizably folk psychological—beliefs about oneself.

Second example: vision. According to many contemporary theories of vision, seeing involves processing information. Some of this information is present in the retinal image; some of it is provided by the visual mechanisms themselves. The information-bearing states postulated by these theories are similar in important ways to the beliefs postulated by folk psychology. For example, both have content and both are involved in rational inferences.

Third example: memory. Folk psychology recognizes that we can store and retrieve information from memory. Scientific psychology also recognizes that fact, although it has gone much further than folk psychology in exploring both the varieties and limitations of human memory. For example, scientific psychology recognizes both *short-* and *long-term* memory, and has explored the relationship between them. Moreover, scientific psychology has discovered that there are quite distinct forms of memory involved in the storage and recall of different sorts of linguistic information. Interestingly, these different sorts of linguistic memory are stored in subtly different areas of the brain. Whilst scientific psychology has made many important discoveries about memory, it's clear that these are discoveries about a process originally identified by folk psychology.

3. *Folk psychology lacks extensive evidential links with the sciences.* One of the striking facts about science is the way scientific theories support each other. Here's my favorite example of this phenomenon. Darwin's theory of natural selection is supported by evidence from a vast range of other scientific endeavors. The theory of continental drift plays an important role in understanding the distribution of species; geology more generally has provided crucial evidence about the age of the Earth. Genetics plays an essential role in explaining how the fittest organisms pass on their genes, and biochemistry has played an essential role in understanding the chemical basis of genetics. Comparative anatomy has helped construct plausible hypotheses about the interrelationships of species, and the physics of isotopes has played a crucial role in dating the ancient remains of animals and plants. The list goes on and on. In each case the theory of natural selection gains support—sometimes a lot; sometimes just a little—from other scientific research.

Any theory which lacks these sorts of connections to other well-established theories is likely to be largely unsupported. According to Churchland, folk psychology is just such a theory, lacking almost entirely significant connections

with other well-established theories. We have therefore further grounds for thinking that folk psychology may indeed be radically false.

Reply. It's hardly surprising that folk psychology lacks a rich network of connections to scientific theories. After all, folk psychology is not a scientific theory. Rather, folk psychology is a collection of platitudes which ordinary people are inclined to accept, and ordinary people are not likely to be sufficiently knowledgeable about science to explore in detail the connections between folk psychology and, say, neurobiology. Moreover, there *are* connections between scientific folk psychology and various other sciences. For example, there is currently a great deal of interest in connecting theories in scientific folk psychology to research in the theory of evolution.

5.4 Anti-eliminativist arguments

So far we have considered three arguments which seek to support eliminativism by debunking folk psychology. In this section I will briefly discuss two anti-eliminativist arguments.

1. *The predictive success of folk psychology.* Eliminativists often draw attention to folk psychology's failings. However, we must not overlook folk psychology's *successes*. A number of theorists—especially the contemporary American philosopher Jerry Fodor—have emphasized how impressive folk psychology is as a predictive tool. Here's an example of a successful folk psychological prediction. My students can predict with considerable reliability where I will be at 10 a.m. next Monday morning: they know that I will be in Lecture Theater North 1. They can do this because they have attributed to me certain folk psychological states. For example, they know that I *believe* that my philosophy of mind lecture starts at 10 a.m. every Monday and is located in Lecture Theater North 1, and they know that I always *like* to get to class on time.

So commonplace are predictions of this kind that we tend to forget how remarkable they are. Notice that predicting where I will be at 10 a.m. next Monday morning is completely beyond the powers of contemporary neurosciences. Even if my brain was subjected to the most rigorous testing currently available, neuroscientists could not predict the movements I will make in five days' time. Nevertheless, my undergraduate students can easily and accurately predict where I will be in five days' time. So when it comes to predicting the movements of human beings, folk psychology completely trumps neuroscience.

A theory which is so predictively successful deserves our respect. Of course, predictive success does not *guarantee* truth. Newton's theories, for example, were staggeringly predictively successful but turned out to be wrong. However, in

general predictive success is evidence in favor of a theory, and folk psychology has predictive success in spades.

2. *The success of scientific folk psychology.* In the previous section we noted that much scientific psychology quantifies over states originally posited by folk psychology. For want of a better term, I called such theories 'scientific folk psychology'. We saw that scientific folk psychology is highly successful at explaining a range of features of our cognitive lives. As we discussed in Section 5.1, the success of a theory gives us good reason to accept the existence of the states over which it quantifies. Since scientific folk psychology is successful, and since it quantifies over folk psychological states, we have good reason to think that those states actually exist.

5.5 Fictionalism

Fictionalism in the philosophy of mind is the doctrine that, whilst strictly speaking there are no mental states, it's extremely useful to pretend that there are. (Fictionalism is also known as 'instrumentalism' since it views the attribution of mental states as having instrumental value—and nothing more.) In this section I'm going to take Daniel Dennett's position as my example of fictionalism. Dennett sometimes objects to being labeled a 'fictionalist'; however, at least some of his writings strongly give the impression that he is one. My apologies to Professor Dennett if I've misrepresented him.

Let's begin by acknowledging just how useful is the ascription of mental states. Following Dennett, we can recognize three 'stances' from which we can predict the behavior of a complex system like a chess-playing computer or a human being: the *physical* stance; the *design* stance; and the *intentional* stance.

1. *The physical stance.* Both chess-playing computers and human beings are physical objects. Setting aside worries about quantum indeterminacy, the behavior of both chess machines and humans can in principle be predicted by treating them as vast assemblages of elementary physical particles, and applying the laws of physics to those particles. Predicting the behavior of a system in this fashion is called 'taking the physical stance'. For all but the simplest systems, the physical stance is unworkable: the number of particles and the complexity of their arrangements makes practical prediction impossible.

2. *The design stance.* Sometimes it is possible to predict the behavior of a complex system by thinking about what it is *supposed to do*. For example, the people who designed my laptop and the software it's running intended it to follow the rule 'When the *p* key is pressed display the letter 'p' on the screen'. Knowing that that's how my laptop is supposed to work, I can predict what will happen when I press the *p* key.

Making predictions about a system's behavior by thinking about what the system is supposed to do is called 'taking the design stance'. Of course, the design stance doesn't always work. If my software has a bug in it, or if I've forgotten to recharge the battery, pressing the *p* key might not result in the letter 'p' appearing on the screen. (I once dropped a cup of coffee on the keyboard of my computer. Thereafter the only key which worked was the *z* key, and it worked whether I pressed it or not!) When the design stance fails to yield accurate predictions we usually retreat to the physical stance. That is, we stop thinking about what the system is supposed to do, and treat it as a physical object which obeys the laws of physics.

Chess-playing computers are artifacts. They are designed by smart people so that they play chess competently. In the case of artifacts it's usually obvious what the system is supposed to do. But what about human beings and other biological systems? What are they 'supposed' to do? At this point Dennett appeals to Darwin's theory of natural selection: biological systems are 'supposed' to do whatever it is that they were selected to do. Eyes, for example, were selected to provide visual information about the organism's environment, so eyes are 'supposed' to see. The inverted commas around 'supposed' are important. If Darwin's right about the evolution of organisms, nobody designed the eye or intended the eye to do anything. Rather, eyes are the outcome of a great many tiny changes to a pre-existing structure. (For a brilliant introduction to the theory of natural selection see Dawkins 1986.) Consequently, if we are being very careful we should say that the design stance predicts what a complex system will do by considering what it was designed *or naturally selected* to do.

3. *The intentional stance.* Sometimes even the design stance is, in practice, unworkable. This happens when the design is too complicated or simply unknown to us. At this point we can adopt the intentional stance. The intentional stance begins with the assumption that the complex system in question is rational—it believes what it should believe and desires what it should desire. For example, the intentional stance assumes that if you are staring at a nearby cow in good light you will come to believe that there is a cow nearby; and that if you need some cash you will desire to go to the bank. (Old joke. Social worker: 'Why do you rob banks?' Criminal: 'Because that's where the money is.')

Now if we assume that the complex system in question is rational, we can predict its behavior. For example, I can predict what you will do when you are driving a car and approach a red light. First, I can assume you believe that the traffic light is red. Second, I can assume you desire to stop at red lights. (By and large, driving through red lights is not a rational form of behavior!) Putting this together, I can predict that you will stop at the red light. And chances are, I'll be right.

In practice attributing mental states to people is indispensable. Applying the physical stance to systems as complex as human beings is very often simply impossible. Moreover, we don't as yet have a complete understanding of what the various neural systems of the human brain were selected for (or 'supposed' to do). Consequently, when it comes to predicting human behavior we usually rely on the attribution of mental states.

How does all this connect with fictionalism? Dennett notes that we can apply the intentional stance to a chess-playing computer, saying things like, 'It wants to save its knight' or, 'It thinks it should get its queen out early'. However, Dennett asserts that if we actually look at the chess-playing program we will find nothing which corresponds to the attributed thoughts. Very roughly, chess-playing computers work by identifying the available moves and assigning each move a number. The number represents the attractiveness of the move, and the computer executes that move which has the highest number. The algorithms which assign the numbers don't contain instructions like, 'Get the queen out early'. Dennett concludes that, whilst attributing beliefs and desires to the computer is very useful—perhaps even unavoidable—it doesn't really have any beliefs and desires. Similarly, whilst attributing beliefs and desires to other people is very useful—perhaps even unavoidable—if you look inside us you quickly realize that there are no such things as beliefs and desires.

Two comments are in order.

1. The argument from the chess-playing computer to fictionalism about mental states in general is too quick. Notice that we're not inclined to take the attribution of mental states to chess-playing computers very seriously. Most of us dismiss talk about what the computer does or does not believe as 'anthropomorphizing'. (To anthropomorphize something is to inappropriately treat it as a human being. Some people anthropomorphize their pet fish.) If we're right not to attribute beliefs and desires to chess-playing computers, then the fact that there's nothing in the program that looks like a belief or a desire isn't surprising. Moreover, it may yet turn out to be the case that the neural correlates of beliefs and desires will be discovered in our heads. At present we simply don't know enough about the brain to rule out finding beliefs and desires inside our skulls.
2. One of the things which Dennett is fond of stressing is that the intentional stance *works*. And surely he's right to this extent: we can very often predict behavior by thinking about the beliefs and desires of the person in question. Now as we saw in Section 5.3, our everyday network of ideas about mental states constitutes a theory—folk psychology. And, as we saw in Section 5.2, other things being equal the predictive success of a theory is good evidence that

the theory is true. It follows that the predictive success of folk psychology is evidence for its truth. In other words, there is a very considerable tension between Dennett's assertion that the intentional stance is so good as to be indispensable, and his claim that mental states are mere fictions. (In this context it's worth recalling Paul Churchland's eliminativist strategy as described in Section 5.3: he didn't praise folk psychology; rather he set out to show that it's a *lousy* theory.)

5.6 Conclusion

We should accept that mental states might not exist—after all, history is full of examples of people believing in things that turned out not to exist. But that's a pretty big 'might'. At present we have little reason to think that mental states don't exist, and consequently we have little reason to endorse either eliminativism or fictionalism.

SUMMARY

(1) Eliminativism is the doctrine that mental states don't exist.

(2) Like eliminativism, fictionalism denies the existence of mental states, but insists that it's very useful to *pretend* that they exist.

(3) Other things being equal, the predictive success of a theory is evidence for its truth.

(4) Taken together, the everyday platitudes about mental states constitute a theory of the mind. That theory is usually called 'folk psychology'.

(5) Eliminativists like Paul Churchland argue that folk psychology is 'radically false' and that consequently we have no reason to accept that there are mental states. However, Churchland's arguments against folk psychology are open to question.

(6) Dennett has identified three 'stances' from which we can predict the behavior of complex systems like chess-playing computers and human beings. Of these, the intentional stance attributes mental states to the system in question on the assumption that the system is rational.

(7) According to Dennett the intentional stance is, in practice, very often the only available means of prediction.

(8) Dennett argues that, whilst we readily attribute mental states to chess-playing computers, there is nothing inside the machine that corresponds to the mental states we have attributed.

(9) Similarly, he holds that whilst attributing mental states to humans is pretty much unavoidable, there are unlikely to be things inside our heads which correspond to mental states.

(10) Dennett's position faces the following difficulty: if mental states are merely fictional, why does attributing them to complex systems work so well?

FURTHER READING

The classic presentation of eliminativism is Paul Churchland's paper 'Eliminativist materialism and the propositional attitudes' (Churchland 1981). This is not merely important and provocative, it's also highly readable. Another important source is Stephen Stich's book, *From Folk Psychology to Cognitive Science* (Stich 1983).

I highly recommend Horgan and Woodward's reply to Churchland, 'Folk psychology is here to stay' (Horgan and Woodward 1985). For a functionalist reply to eliminativism see Jackson and Pettit 1993. Jerry Fodor brilliantly defends folk psychology in his book *Psychosemantics* (Fodor 1987). Chapter 1 is especially recommended.

The idea of a scientific research program is due to Imre Lakatos. See, for example, Lakatos and Zahar 1978. For a good discussion of Lakatos's views see Newton-Smith 1981: Ch. 4.

Dennett's most important papers are 'Intentional Systems' (Dennett 1971) and 'True Believers' (Dennett 1975). Dennett sometimes objects to being labeled a 'fictionalist'; however, you could be forgiven for thinking he is one. He discusses his attitude to realism about mental states in his paper 'Reflections: Real patterns, deeper facts, and empty questions' (Dennett 1987*b*). Fodor briefly raises the issue of why folk psychology works if it's actually false in his 1990*a*. (My guess is that he's not the only one to air this worry.)

Braddon-Mitchell and Jackson 1996: Ch. 13 and Sterelny 1990: Ch. 7 are both excellent secondary sources on eliminativism. Braddon-Mitchell and Jackson 1996: Ch. 9 is also good on the intentional stance and fictionalism.

TUTORIAL QUESTIONS

(1) What is eliminativism?

(2) What is folk psychology?

(3) Sketch Churchland's reasons for thinking that folk psychology is radically false. Do you think that his reasons are good ones?

(4) Discuss the following argument. 'Churchland tells us that there are no such things as beliefs. In other words he *believes* that there are no such things as beliefs. But that's a contradiction. So eliminativism is false.'

(5) Describe Dennett's three stances.

(6) Why does Dennett think that the chess-playing computer does not really have beliefs and desires?

(7) 'The predictive success of folk psychology gives us good reason to reject Dennett's fictionalism.' Discuss.